Titre: Title:	Forestogram: Biclustering Visualization Framework with Applications in Public Transport and Bioinformatics
Auteur: Author:	Mohammad Sajjad Ghaemi
Date:	2017
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ghaemi, M. S. (2017). Forestogram: Biclustering Visualization Framework with Applications in Public Transport and Bioinformatics [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/2904/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/2904/
Directeurs de recherche: Advisors:	Bruno Agard, & Vahid Partovi Nia
Programme: Program:	Doctorat en mathématiques de l'ingénieur

UNIVERSITÉ DE MONTRÉAL

FORESTOGRAM: BICLUSTERING VISUALIZATION FRAMEWORK WITH APPLICATIONS IN PUBLIC TRANSPORT AND BIOINFORMATICS

MOHAMMAD SAJJAD GHAEMI
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(MATHÉMATIQUES DE L'INGÉNIEUR)
DÉCEMBRE 2017

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

FORESTOGRAM: BICLUSTERING VISUALIZATION FRAMEWORK WITH APPLICATIONS IN PUBLIC TRANSPORT AND BIOINFORMATICS

présentée par : GHAEMI Mohammad Sajjad
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de :

M. ADJENGUE Luc-Désiré, Ph. D., président
M. AGARD Bruno, Doctorat, membre et directeur de recherche
M. PARTOVI NIA Vahid, Doctorat, membre et codirecteur de recherche
Mme MORENCY Catherine, Ph. D., membre
Mme LEFEBVRE Geneviève, Ph. D., membre externe

RÉSUMÉ

Dans de nombreux problèmes d'analyse de données, les données sont exprimées dans une matrice avec les sujets en ligne et les attributs en colonne. Les méthodes de segmentations traditionnelles visent à regrouper les sujets (lignes), selon des critères de similitude entre ces sujets. Le but est de constituer des groupes de sujets (lignes) qui partagent un certain degré de ressemblance. Les groupes obtenus permettent de garantir que les sujets partagent des similitudes dans leurs attributs (colonnes), il n'y a cependant aucune garantie sur ce qui se passe au niveau des attributs (les colonnes). Dans certaines applications, un regroupement simultané des lignes et des colonnes appelé biclustering de la matrice de données peut être souhaité. Pour cela, nous concevons et développons un nouveau cadre appelé Forestogram, qui permet le calcul de ce regroupement simultané des lignes et des colonnes (biclusters) dans un mode hiérarchique. Le regroupement simultané des lignes et des colonnes de manière hiérarchique peut aider les praticiens à mieux comprendre comment les groupes évoluent avec des propriétés théoriques intéressantes. Forestogram, le nouvel outil de calcul et de visualisation proposé, pourrait être considéré comme une extension 3D du dendrogramme, avec une fusion orthogonale étendue. Chaque bicluster est constitué d'un groupe de lignes (ou de sujets) qui déplie un schéma fortement corrélé avec le groupe de colonnes (ou attributs) correspondantes. Cependant, au lieu d'effectuer un clustering bidirectionnel indépendamment de chaque côté, nous proposons un algorithme de biclustering hiérarchique qui prend les lignes et les colonnes en même temps pour déterminer les biclusters. De plus, nous développons un critère d'information basé sur un modèle qui fournit un nombre estimé de biclusters à travers un ensemble de configurations hiérarchiques au sein du forestogramme sous des hypothèses légères. Nous étudions le cadre suggéré dans deux perspectives appliquées différentes, l'une dans le domaine du transport en commun, l'autre dans le domaine de la bioinformatique.

En premier lieu, nous étudions le comportement des usagers dans le transport en commun à partir de deux informations distinctes, les données temporelles et les coordonnées spatiales recueillies à partir des données de transaction de la carte à puce des usagers. Dans de nombreuses villes, les sociétés de transport en commun du monde entier utilisent un système de carte à puce pour gérer la perception des tarifs. L'analyse de cette information fournit un aperçu complet de l'influence de l'utilisateur dans le réseau de transport en commun interactif. À cet égard, l'analyse des données temporelles, décrivant l'heure d'entrée dans le réseau de transport en commun est considérée comme la composante la plus importante des données recueillies à partir des cartes à puce. Les techniques classiques de segmentation, basées sur la distance, ne sont pas appropriées pour analyser les données temporelles. Une nouvelle

projection intuitive est suggérée pour conserver le modèle de données horodatées. Ceci est introduit dans la méthode suggérée pour découvrir le modèle temporel comportemental des utilisateurs. Cette projection conserve la distance temporelle entre toute paire arbitraire de données horodatées avec une visualisation significative. Par conséquent, cette information est introduite dans un algorithme de classification hiérarchique en tant que méthode de segmentation de données pour découvrir le modèle des utilisateurs. Ensuite, l'heure d'utilisation est prise en compte comme une variable latente pour rendre la métrique euclidienne appropriée dans l'extraction du motif spatial à travers notre forestogramme.

Comme deuxième application, le forestogramme est testé sur un ensemble de données multiomiques combinées à partir de différentes mesures biologiques pour étudier comment l'état de santé des patientes et les modalités biologiques correspondantes évoluent hiérarchiquement au cours du terme de la grossesse, dans chaque bicluster. Le maintien de la grossesse repose sur un équilibre finement équilibré entre la tolérance à l'allogreffe fœtale et la protection mécanismes contre les agents pathogènes envahissants. Malgré l'impact bien établi du développement pendant les premiers mois de la grossesse sur les résultats à long terme, les interactions entre les divers mécanismes biologiques qui régissent la progression de la grossesse n'ont pas été étudiées en détail. Démontrer la chronologie de ces adaptations à la grossesse à terme fournit le cadre pour de futures études examinant les déviations impliquées dans les pathologies liées à la grossesse, y compris la naissance prématurée et la prééclampsie. Nous effectuons une analyse multi-physique de 51 échantillons de 17 femmes enceintes, livrant à terme. Les ensembles de données comprennent des mesures de l'immunome, du transcriptome, du microbiome, du protéome et du métabolome d'échantillons obtenus simultanément chez les mêmes patients. La modélisation prédictive multivariée utilisant l'algorithme Elastic Net est utilisée pour mesurer la capacité de chaque ensemble de données à prédire l'âge gestationnel. En utilisant la généralisation empilée, ces ensembles de données sont combinés en un seul modèle. Ce modèle augmente non seulement significativement le pouvoir prédictif en combinant tous les ensembles de données, mais révèle également de nouvelles interactions entre différentes modalités biologiques. En outre, notre forestogramme suggéré est une autre ligne directrice avec l'âge gestationnel au moment de l'échantillonnage qui fournit un modèle non supervisé pour montrer combien d'informations supervisées sont nécessaires pour chaque trimestre pour caractériser les changements induits par la grossesse dans Microbiome, Transcriptome, Génome, Exposome et Immunome réponses efficacement.

ABSTRACT

In many statistical modeling problems data are expressed in a matrix with subjects in row and attributes in column. In this regard, simultaneous grouping of rows and columns known as biclustering of the data matrix is desired. We design and develop a new framework called Forestogram, with the aim of fast computational and hierarchical illustration of biclusters. Often in practical data analysis, we deal with a two-dimensional object known as the data matrix, where observations are expressed as samples (or subjects) in rows, and attributes (or features) in columns. Thus, simultaneous grouping of rows and columns in a hierarchical manner helps practitioners better understanding how clusters evolve. Forestogram, a novel computational and visualization tool, could be thought of as a 3D expansion of dendrogram, with extended orthogonal merge. Each bicluster consists of group of rows (or samples) that unfolds a highly-correlated schema with their corresponding group of columns (or attributes). However, instead of performing two-way clustering independently on each side, we propose a hierarchical biclustering algorithm which takes rows and columns at the same time to determine the biclusters. Furthermore, we develop a model-based information criterion which provides an estimated number of biclusters through a set of hierarchical configurations within the forestogram under mild assumptions. We study the suggested framework in two different applied perspectives, one in public transit domain, another one in bioinformatics field.

First, we investigate the users' behavior in public transit based on two distinct information, temporal data and spatial coordinates gathered from smart card. In many cities, worldwide public transit companies use smart card system to manage fare collection. Analysis of this information provides a comprehensive insight of user's influence in the interactive public transit network. In this regard, analysis of temporal data, describing the time of entering to the public transit network is considered as the most substantial component of the data gathered from the smart cards. Classical distance-based techniques are not always suitable to analyze this time series data. A novel projection with intuitive visual map from higher dimension into a three-dimensional clock-like space is suggested to reveal the underlying temporal pattern of public transit users. This projection retains the temporal distance between any arbitrary pair of time-stamped data with meaningful visualization. Consequently, this information is fed into a hierarchical clustering algorithm as a method of data segmentation to discover the pattern of users. Then, the time of the usage is taken as a latent variable into account to make the Euclidean metric appropriate for extracting the spatial pattern through our forestogram.

As a second application, forestogram is tested on a multiomics dataset combined from dif-

ferent biological measurements to study how patients and corresponding biological modalities evolve hierarchically in each bicluster over the term of pregnancy. The maintenance of pregnancy relies on a finely-tuned balance between tolerance to the fetal allograft and protective mechanisms against invading pathogens. Despite the well-established impact of development during the early months of pregnancy on long-term outcomes, the interactions between various biological mechanisms that govern the progression of pregnancy have not been studied in details. Demonstrating the chronology of these adaptations to term pregnancy provides the framework for future studies examining deviations implicated in pregnancy-related pathologies including preterm birth and preeclampsia. We perform a multiomics analysis of 51 samples from 17 pregnant women, delivering at term. The datasets include measurements from the immunome, transcriptome, microbiome, proteome, and metabolome of samples obtained simultaneously from the same patients. Multivariate predictive modeling using the Elastic Net algorithm is used to measure the ability of each dataset to predict gestational age. Using stacked generalization, these datasets are combined into a single model. This model not only significantly increases the predictive power by combining all datasets, but also reveals novel interactions between different biological modalities. Furthermore, our suggested forestogram is another guideline along with the gestational age at time of sampling that provides an unsupervised model to show how much supervised information is necessary for each trimester to characterize the pregnancy-induced changes in Microbiome, Transcriptome, Genome, Exposome, and Immunome responses effectively.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF NOTATIONS AND ACRONYMS

NOTATIONS:

$\mathcal{C}_i$	biclusters i
n	number of subjects, or number of rows
p	number of attributes, or number of columns
$\mathbf{Y}_{n\times p}$	$n \times p$ data matrix
$\check{\mathbf{Y}}$	submatrix of $\mathbf{Y}$, i.e. $\check{\mathbf{Y}} \subset \mathbf{Y}_{n\times p}$
$\bar{y}(.)$	mean of cluster
$\tilde{y}(.)$	median of cluster
σ^2	common within variance
ϕ	between to within variance ratio
s^2	pooled variance
$\mathfrak{M}(.)$	margin of bicluster
$\mathfrak{D}(.)$	diameter of bicluster
$D(.,.)$	dissimilarity measure between a pair of biclusters
S_i	spatial usage of card i
X_i	binary temporal usage of card i
r_i	temporal boarding radius of card i
θ_i	temporal boarding angle of card i
z_i	temporal boarding variance of card i
$\odot$	Hadamard (elementwise) product operator
$\|\|.\|\|$	Euclidean norm
$\|c\|$	absolute value of the scalar c
$\|\mathcal{C}\|$	cardinality of the set $\mathcal{C}$

ACRONYMS:

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
DBSCAN	Density-Based Spatial Clustering of Application with Noise
DTW	Dynamic Time Warping
EDM	Euclidean Distance Matrix
FORIC	FORest Information Criterion
GIS	Geographic Information System
ISFCS	Integrated Smart Card Fare Collection System

MST	Minimum Spanning Tree
NMF	Nonnegative Matrix Factorization
RGCCA	Regularized generalized canonical correlation analysis
SCP	Semi-Circle Projection
SCFCS	Smart Card Fare Collection System
STO	Société de Transport de l'Outaouais

CHAPTER 1 INTRODUCTION

The rapid growth of data according to the progress of sensors and storage technologies has been emerging in several different areas (Jain, 2010). Various sources can generate these data, from the Internet search, digital videos, imaging and biological sequences to smart card data used in public transit. Therefore many researchers and scientists from miscellaneous fields such as mathematics, statistics, computer science, urban computing and planning, management, business, civil engineering, industrial engineering, Geographic Information System (GIS), and biology have encouraged to concentrate on finding methods for grouping a set of data (Jain, 2010; Everitt *et al.*, 2011). The main purpose of extracting similar groups of data without label information is to discover knowledge and interpret the high volume data before deep analyzing the fine-grained components that are hidden in the underlying data. In the most simplest way, clustering aims to ensure giving a coherent and complementary big picture of a complex dataset to figure out what one can do with the data where everything is intricate. In marketing, similar group of customers with similar commercial habits or demographic can be found by clustering (Murray *et al.*, 2017; Huang *et al.*, 2007; Punj and Stewart, 1983). In biology, grouping similar diseases, genes or phenotypes according to the different level of measurements is easily carried out by clustering (Nugent and Meila, 2010; Ben-Dor *et al.*, 1999; Eisen *et al.*, 1998; Eren *et al.*, 2013). In public transport and city planning, clustering is used to identify similar users, passengers profiling, station grouping, and infrastructure development (Pelletier *et al.*, 2011; Carel and Alquier, 2017; Nin *et al.*, 2013; Galba *et al.*, 2013; Vos and Witlox, 2013). Clustering has many other interesting applications in digital domain, such as document retrieval, image segmentation, recommender systems, search engine, social networks, etc. (Orzechowski and Boryczko, 2016; Zamir and Etzioni, 1998; Huang, 2008; Pal and Pal, 1993). The formation of coherent data as a unit of cluster together could be carried out according to a *measure of similarity* which reflects the relationships among the data. Thus the increase of data generation, in both capacity and diversity, demands fundamental improvement in methodological and algorithmic methods toward spontaneously realizing, processing and extracting the patterns underlying the data.

One of the most important and challenging tasks in machine learning, statistics and generally data analysis is grouping similar objects together (Kleinberg, 2003; von Luxburg *et al.*, 2012). In recent years, this task known as clustering, has arisen as a progressively important research topic both in theory and application such as pattern recognition, data mining, bioinformatics, computer vision, social network, etc. (Eisen *et al.*, 1998; Madeira and Oliveira, 2004; Orzechowski and Boryczko, 2016; Tu and Honavar, 2008; Jain, 2010). Clustering

analysis, due to the absence of the label information, is called *unsupervised learning*. Unlike the other existing problems in this field, such as, classification or regression, in the study of clustering analysis, there is no a priori knowledge available to identify the category label information for the given data.

Unlike confirmatory methods that deal with validating the given assumptions of the model to the data, the purpose of exploratory methods, is to discover and extract groups of data, known as *data clusters*, into interpretable and meaningful information to the specialists (von Luxburg *et al.*, 2012). Clustering is an exceptionally tough combinatorial problem that belongs to the NP-hard class of computational complexity problems (Ackerman and Ben-david, 2009). Indeed, it was shown that there is no clustering algorithm that is able to preserve certain properties for data clustering (Kleinberg, 2003). In this regime, clustering is more likely considered as an art rather than science (von Luxburg *et al.*, 2012). To cope with this difficulty, quite many different techniques were already suggested in plenty of contexts by numerous researchers which demonstrate the broad necessity and appeal to the exploratory data analysis problem (Jain, 2010).

Usually, there is no right or wrong clusters. Because evaluating a clustering algorithm depends on why user does clustering and how the result of clustering can be used (von Luxburg *et al.*, 2012). In the case of low-dimensional data (ideally less than 4) by plotting the data, we can distinguish the clusters in the data visually. However, recently, the advent of high-dimensional data demonstrates the need for devising new algorithms to reveal the clusters by exploring the data. Therefore, appropriate clustering criterion can be deployed to extract suitable clustering assignment from the data to meet users' demand. Clustering algorithms are categorized into two groups: hierarchical and partitional (Jain, 2010). In the former category, hierarchical methods try to find nested clusters recursively, while in partitional approaches data are split into a non-overlapping division of subsets without any nested procedure. Hierarchical methods are being used widely in applied projects, especially for public transit (Patnaik *et al.*, 2016; Wang and Yu, 2010) and bioinformatics (Pontes *et al.*, 2015; Madeira and Oliveira, 2004). In order to address these applications successfully, one of the concerning questions that should be answered significantly is the estimation for the number of clusters. Consequently, the cutting point on the dendrogram at certain height to illustrate a set of major clusters existing in the hierarchical configuration has been a well-known problem for decades. The majority of algorithms for this regard can be divided into distance-based or model-based methods (Stahl and Sallis, 2012; Oh and Raftery, 2007; Farrar, 2006; Tan *et al.*, 2006; Izenman, 2008). Distance-based techniques are easy to understand and simple to implement. On the contrary, model-based approaches are flexible and adapt to data pattern, but are counter intuitive to implement. However some methods are developed for distance-

based methods using cross validation (Tibshirani *et al.*, 2001), or for model-based methods using statistical asymptotic (Claeskens and Hjort, 2008). Our suggestion is to apply a simple Bayesian hierarchical model that allows to apply the ratio of posterior predictive (Kass and Raftery, 1995), as the standard model selection criteria. Hence, we develop a *Model-based Clustering* that assumes a statistical model for clustering the data based on hierarchical settings with promising results in real world applications (McLachlan *et al.*, 2004).

Hierarchical clustering is a breakthrough in clustering, because of producing a visual guide in the form of a binary tree, known as dendrogram. In addition, it requires little prior knowledge, except for a dissimilarity measure (Johnson, 1967; Sokal and Sneath, 1963). The dissimilarity measure is a positive semi-definite symmetric mapping of pairs of groups onto the set of real numbers (Murtagh and Legendre, 2014). This measure, however, may not satisfy the triangle inequality unlike the *distance* (Murtagh and Legendre, 2014). Hierarchical algorithms require a dissimilarity measure to merge clusters in order to build a nested structure of clusters. The common dissimilarities include single linkage (or nearest neighbors), complete linkage (or farthest neighbors), average linkage, and centroid linkage (Murtagh and Legendre, 2014) among others. There are two variants of hierarchical clustering methods depending on the direction of the construction of the nested groups. *Agglomerative clustering* starts with every observation as a singleton and consequently merges the closest clusters to end up with all data in one cluster (Johnson, 1967). *Divisive algorithms*, on the contrary, starts with all data in one cluster and splits the clusters until finishing with all singletons (Everitt *et al.*, 2011).

Clustering could be applied on public transit domain where everyday, thousands of people travel (de Oña and de Oña, 2015; Ghasemzadeh *et al.*, 2014; Hasan *et al.*, 2012; Fuse *et al.*, 2012). Each time a smart card is tapped on the card reader, plenty of information is gathered that possibly could lead analysts, engineers, managers and strategists to excavate, design, decide, and plan more effectively based on the users behavior in this network (Dou *et al.*, 2015; Park *et al.*, 2008; Kurauchi and Schmöcker, 2017; Pelletier *et al.*, 2011). Investigating users behavior according to the data is a nontrivial task. It requires sophisticated mathematical, statistical, data mining and machine learning techniques to exploit the hidden patterns of the stored data. Such data is dynamic and rigorously increasing because of population growth and development of infrastructure. Moreover, affordable cost of public transit in comparison to private car, especially, in the large cities, metropolitan areas and their suburb increases the daily usage of this network. In this regard, we propose new methods of spatial-temporal data analysis to investigate patterns of user's behavior in public transit.

The importance of the public transportation and its influence in the real life of many people in large cities around the world rises a new family of problems that is not confined

into a particular branch of science (Weisbrod and Reno, 2009). Hence, usage of the smart card data creates the opportunity for several different researchers from diverse disciplines e.g. data mining, machine learning, urban computing and planning, management, business, civil engineering, industrial engineering, statistics, mathematical engineering, GIS, etc. to outreach and extend their methods to analyze the data for the public transport authorities (Weisbrod and Reno, 2009; Gallotti and Barthelemy, 2015; Fuse *et al.*, 2012; Ortega-Tong, 2013; Lathia *et al.*, 2010).

Despite extensive researches have been done on public transportation domain, various obstacles have been arisen for specific purposes which require particular approaches to address them. In this study, a recent problem of clustering the similar users is introducing according to the spatial-temporal data gathered from smart cards to analyze the user's behavior in the public transport network.

Smart card data, contains worthwhile digital information of daily locations visited at certain period of a large number of individuals. The wealth of collected data down to a single time, and location resolution creates fundamental challenges that require a mix of analytical, algorithmic, and statistical techniques (Pelletier *et al.*, 2011). Beside other sources of information such as mobile phone, GPS tracker vehicle, e.g. bike, car, motorcycle, credit card transactions, social network, and many other sources of information gathering, smart card data is the best promising source of users digital information (Hasan *et al.*, 2012). Thus this helpful information could be utilized to characterize and model urban mobility patterns (Hasan *et al.*, 2012). Other useful information such as travel time and number of passengers for the sake of congestion analysis and planning improvement, could be possibly extracted as well (Fuse *et al.*, 2012).

From a different perspective, high-throughput technologies such as cell-free RNA, plasma luminex, serum luminex, immune system, metabolomics, and plasma somalogic have been recently studied for different biological researches e.g. aging, recovering from surgery, stroke, pregnancy, cancer, etc. in order to provide statistical predictive models for diagnosis, prognosis and therapy (Clarke *et al.*, 2008; Yau *et al.*, 2016; Dey *et al.*, 2017; Ji and Liu, 2010). Every single biological dataset consists of hundreds or thousands of highly correlated measurements for each sample so that extracting meaningful biological information from these high-dimensional datasets through statistical techniques delivers tremendous insights for biological investigators (Schwenk *et al.*, 2010; Bendall *et al.*, 2011; Kang *et al.*, 2008; Sreekumar *et al.*, 2009; Miyagi *et al.*, 2010; Katz *et al.*, 2016; Romero *et al.*, 2017; Clarke *et al.*, 2008). Devising, designing and implementing statistical data analysis methods make the biologists better comprehending the role of certain group of genes, proteins, and many other biological measurements intuitively to address the critical questions in drug development, treatment

strategy, gene signaling pathway, etc (Miller *et al.*, 2008). Toward the goal of making the sense of data for biologists, visualization plays a central role especially for displaying hierarchical structures. Hierarchical clustering and its applications are familiar to many biologists so that pairwise relationships between data points are represented by a rooted binary tree. This kind of clustering is widely used for evolutionary analysis of sequence history e.g. phylogenetic sequencing, gene expression patterns, DNA microarray expression, etc. (Eisen *et al.*, 1998; Clarke *et al.*, 2008; Miller *et al.*, 2008; Xu and Wunsch, 2005). Intuitive illustration of groups without complicated assumptions about the intrinsic of the data distribution, along with the effective computational complexity make the hierarchical approach the first choice for analyzing biological data (Eisen *et al.*, 1998).

The main goal of this thesis is the development of forestogram framework for hierarchical biclustering with the applications in public transit and bioinformatics. In practical applications, a method that is easy to understand by people is desirable. Hierarchical clustering is an intuitive method of clustering for many people in industrial engineering and bioinformatics. Additionally, binary tree representation known as dendrogram is easy to elaborate the result such that the result interpretation is easy enough to provide a descriptive summary of the data. In public transit domain, we often deal with two types of distinct information, temporal and spatial. For the temporal data that are similar to time series data, the hierarchical clustering techniques are not suitable because off-the-shelf distance metrics are not designed for binary vectors. To this end, we first suggest a projection technique to map a long binary vector of temporal usage into three dimensional space which retains the proximity of pairwise similarity. For the next step, we take the temporal information as a latent variable to extract the spatial-temporal patterns from the data such that Euclidean metric becomes feasible because of geodesic property of GPS location history. In the context of multiomics analysis of gestational age prediction, we suggest to analyze each dataset separately to show how each biological measurement is influencing the term of pregnancy. For the integrative model, we suggest the stacked generalization and forestogram approaches for supervised and unsupervised integrative data analysis.

In Chapter 4, we address the biclustering problem in general and as an extension to the idea of model-based hierarchical clustering in particular where we have a matrix that consists of rows and columns such that grouping of both sides is the case of interest. This is a variant of normal grouping of data but instead of only similar groups of samples, a block of submatrix with subjects that are highly correlated with features forms the biclusters. Analogous to the conventional linkages for hierarchical clustering, we suggest a bilinkage dissimilarity measure for constructing a hierarchical setting for these biclusters. Consequently, we also develop forestogram visualization technique similar to dendrogram with one extra

dimension to emphasize the direction of the merge when switches from row to column or vice-versa. Then we elaborate how to find the number of representative biclusters on the forestogram by making a subtle connection from hierarchical fashion to the model based setting. This way, Bayesian viewpoint helps us estimating this number by deploying *FORIC* that is a kind of information criterion trick motivated by *Bayesian Information Criterion* (BIC). Then this framework is tested on synthesized, plus real world instances such as public transit, and biological datasets. Promising results from simulation and empirical studies turn out that forestogram outperforms almost all rival techniques comparing with histogram and a number of similar methods.

In Chapter 5, first we introduce a temporal projection that maps a high-dimensional binary vector corresponding to the hourly usage of public transit into a space of three dimension with a semi-circle shape. This ad-hoc projection is exclusively designed to mimic the clock as a clue for temporal data with a number of interesting mathematical properties that retains the similar users close by. Despite many metrics defined for computing the distance between an arbitrary pair of data, such as Euclidean, Manhattan, Hamming, etc. none of them satisfies the relations for temporal binary vectors. To this end, our suggested projection meets certain properties that are suitable for analyzing the temporal public transit data. One of the most interesting advantages is the visualization scheme so that makes it easy for transport analysts to better understand how a hierarchical clustering model works on top of this projection. Furthermore, adding the spatial data which represents the geographical location history for associated time-points, enables the modified version of forestogram to specify the spatial-temporal user behavior through the standard Euclidean metric. These methodologies are inspired by the Société de transport de l'Outaouais data with successful results that are present in the experiment section.

Chapter 6, is devoted to the integrative clock of human pregnancy in three trimesters before delivery and postpartum with seven multiomics measurements to investigate the level of changes in these datasets over the term of pregnancy. The goal of this study is to find significant features that are correlated with gestational age in three trimesters before delivery. We develop an algorithm which determines the significant correlated features leading the gestational age on each dataset apart at first step, then an integrative step is implemented in two different levels, features pool and predictions stack. In the feature level, all datasets together make a holistic model for all patients, while the stack generalization outlook takes the independent predictions from each single dataset as a new feature to predict the gestational age. Moreover, due to the ethnicity and particular specifications of women, different features could affect certain patients toward predicting the gestational age. We compare the performance of forestogram as an unsupervised biclustering technique with the elastic net

(Zou and Hastie, 2005) as a supervised variable selection regression method to figure out how much supervised information is necessary toward predicting the trimesters of gestational age in addition to the important features that are correlated with the clock of pregnancy.

CHAPTER 2 LITERATURE REVIEW

2.1 Hierarchical Clustering

The number of ways for partitioning a set to examine all possible clustering increases exponentially in terms of the number of data points and grouping. In this regard, hierarchical clustering algorithms are designed to discover the underlying clusters in a given dataset efficiently. Looking for a reasonable setting of clusters without refining all possible combinatorial assignments is made possible through hierarchical algorithms (Rencher, 1998).

Agglomerative hierarchical clustering consists of a bottom-up sequential process so that the number of clusters shrinks by starting from all singleton clusters whereas the volume of clusters grows gradually by ending up in one cluster surrounding all. In contrast, top-down divisive approach is the opposite viewpoint where a single cluster contains all data points in the beginning then splits into two groups in the next step. The end result of the divisive method is exactly the same as the start point of the agglomerative algorithm (Everitt *et al.*, 2011). For better understanding the mechanism of constructing a dendrogram as a visualization technique for hierarchical clustering, an example is shown in Figure 2.2 where two separable classes of univariate data are given in Figure 2.1.

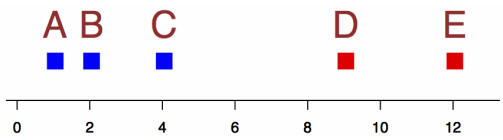


Figure 2.1 A univariate example of 5 data points.

In Figure 2.2 a simple example is shown where in the beginning, we have 5 clusters such that every single data point constitutes the finest clusters (each data point is a singleton cluster) see Figure 2.2(a). As it could be seen in Figure 2.2(b), data A and B are the most proximate pair of data that are agglomeratively combined together to create a new cluster at this step. Then, data point C joins the newly formed cluster, i.e. (A, B) as the closest cluster among the remaining ones in Figure 2.2(c). In the next step in Figure 2.2(d), (D, E) pair are merged together to form a new cluster. Then eventually, after combination of (D, E) to $((A, B), C)$ we get the coarsest cluster to the given set (all data points in one cluster).

Bottom-up approach is easy to form the hierarchy of clusters based on a dissimilarity measure in comparison to top-down method which requires a priori knowledge about the structure and shape of clusters. Having little prior knowledge, except for the dissimilarity

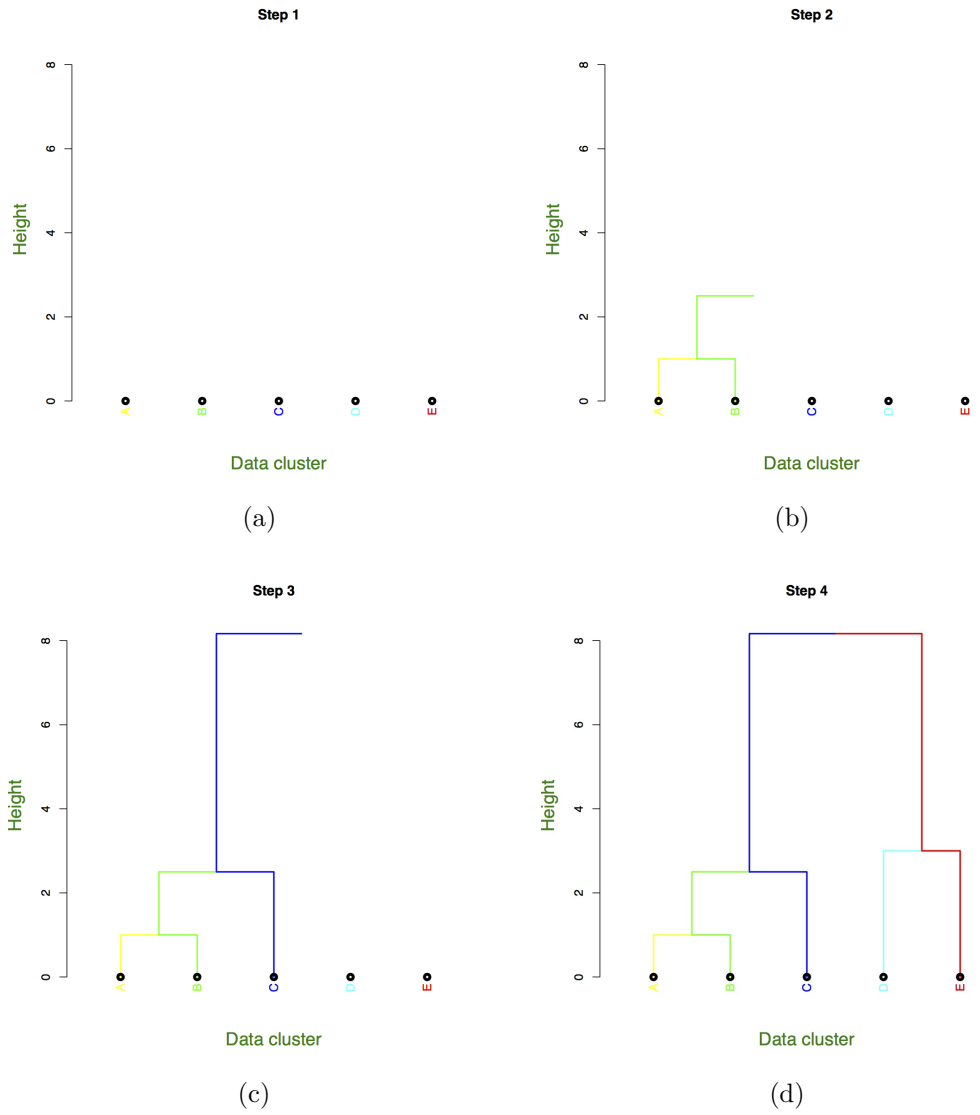


Figure 2.2 Step by step demonstration of agglomerative hierarchical clustering.

measure is one of the advantages of hierarchical clustering algorithms. The dissimilarity measure is a positive semi-definite symmetric mapping of pairs of groups onto the set of real numbers. This measure, however, may not satisfy the triangle inequality unlike the distance. The common dissimilarity measures include, single linkage or nearest neighbours (Florek *et al.*, 1951; Sneath, 1957; Johnson, 1967), complete linkage or farthest neighbours (Sørensen, 1948), average linkage (Sokal, 1958), centroid linkage (Eisen *et al.*, 1998), median linkage, and Ward's linkage or minimum variance (Murtagh and Legendre, 2014).

In Chapter 4 we extend the definition of dissimilarity measure in the form of bilinkage such that the hierarchical biclustering can be constructed to illustrate the nested structure

of block-clusters by the associated forestogram. Apart from a pair of biclusters with dissimilarity measure that should be merged together, direction of the merge is also necessary for forestogram to show how a pair of block-clusters is correlated.

Here we briefly introduce the existing linkages for hierarchical clustering so that bilinkage can be define respectively. In the following notation, $\mathbf{y}_i \in \mathbb{R}^p$ represents a data point belonging to a certain cluster, $d(\mathbf{y}_i, \mathbf{y}_j)$ is the Euclidean distance, then the squared Euclidean distance using norm $||.||$ can be defined as $d^2(\mathbf{y}_i, \mathbf{y}_j) = ||\mathbf{y}_i - \mathbf{y}_j||^2 = \sum_{k=1}^{p}(\mathbf{y}_{ik} - \mathbf{y}_{jk})^2$, and $||\mathcal{C}_i||$ refers to the number of data points in the cluster $\mathcal{C}_i$. Figure 2.3 elaborates the role of each linkage on a random dataset.

2.1.1 Single Linkage

Early single linkage, also known as the nearest neighbor clustering, is one of the oldest and most famous of the hierarchical techniques that is developed by (Florek *et al.*, 1951; Sneath, 1957; Johnson, 1967) which assumes no cluster shape to produce more dense, and chain-like clusters. Single linkage tends to merge close data points or singleton clusters together due to the early merge of two partitions; this undesired property is known as chaining effect see Figure 2.3(a). In other words, a chain of singleton clusters can be extended for long distances against the general form of the cluster. In the single linkage, the distance between two disjoint clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as,

$$D_{\text{single}}(\mathcal{C}_1, \mathcal{C}_2) = \min_{\mathbf{y}_i \in \mathcal{C}_1, \mathbf{y}_j \in \mathcal{C}_2} d(\mathbf{y}_i, \mathbf{y}_j)$$

2.1.2 Complete Linkage

Complete linkage, also known as furthest neighbor or maximum method, is initiated by Sørensen (1948) that roughly produces clusters with almost equal diameters. Complete linkage suffers from the opposite drawback of single linkage problem. If data contains outliers the complete linkage may not combine two proximate clusters in the appropriate order of merge in the hierarchical path see Figure 2.3(b). In the complete linkage, the distance between two disjoint clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as,

$$D_{\text{complete}}(\mathcal{C}_1, \mathcal{C}_2) = \max_{\mathbf{y}_i \in \mathcal{C}_1, \mathbf{y}_j \in \mathcal{C}_2} d(\mathbf{y}_i, \mathbf{y}_j)$$

2.1.3 Average Linkage

Average linkage, also called the weighted pair-group method, is suggested as a trivial compromise between the single linkage and the complete linkage in Sokal (1958). Average

linkage prefers combining clusters with small variances that can be figured out as clusters with the same variance (Sokal, 1958). In the average linkage, the distance between two disjoint clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as,

$$D_{\text{average}}(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{|\mathcal{C}_1|\ |\mathcal{C}_2|} \sum_{\mathbf{y}_i \in \mathcal{C}_1} \sum_{\mathbf{y}_j \in \mathcal{C}_2} d(\mathbf{y}_i, \mathbf{y}_j)$$

2.1.4 Ward Linkage

Ward's linkage minimizes the total within-cluster variance by a weighted squared distance between cluster centers. Ward's method merges clusters to maximize the likelihood at each iteration where spherical covariance matrices is deemed see Figure 2.3(d). Merging pair of clusters with few samples is what preferred by Ward's method to generate balanced size clusters (Milligan, 1980). In the Ward's linkage, the distance between two disjoint clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as,

$$\begin{aligned} D_{\text{Ward}}(\mathcal{C}_1, \mathcal{C}_2) &= \sum_{\mathbf{y}_i \in \mathcal{C}_1 \cup \mathcal{C}_2} ||\mathbf{y}_i - \bar{\mathbf{y}}(\mathcal{C}_1 \cup \mathcal{C}_2)||^2 - \sum_{\mathbf{y}_i \in \mathcal{C}_1} ||\mathbf{y}_i - \bar{\mathbf{y}}(\mathcal{C}_1)||^2 - \sum_{\mathbf{y}_i \in \mathcal{C}_2} ||\mathbf{y}_i - \bar{\mathbf{y}}(\mathcal{C}_2)||^2 \\ &= \frac{|\mathcal{C}_1|\ |\mathcal{C}_2|}{|\mathcal{C}_1| + |\mathcal{C}_2|} ||\bar{\mathbf{y}}(\mathcal{C}_1) - \bar{\mathbf{y}}(\mathcal{C}_2)||^2 \end{aligned}$$

where $\bar{\mathbf{y}}(\mathcal{C}_1)$ and $\bar{\mathbf{y}}(\mathcal{C}_2)$ are the mean vectors of clusters $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

2.1.5 Centroid Linkage

Centroid linkage, also referred to as the unweighted pair-group centroid method, simply minimizes the squared Euclidean distance between cluster means and is less sensitive to outliers in comparison to the other linkages (Milligan, 1980). In the centroid linkage, the distance between two disjoint clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as,

$$D_{\text{centroid}}(\mathcal{C}_1, \mathcal{C}_2) = ||\bar{\mathbf{y}}(\mathcal{C}_1) - \bar{\mathbf{y}}(\mathcal{C}_2)||^2$$

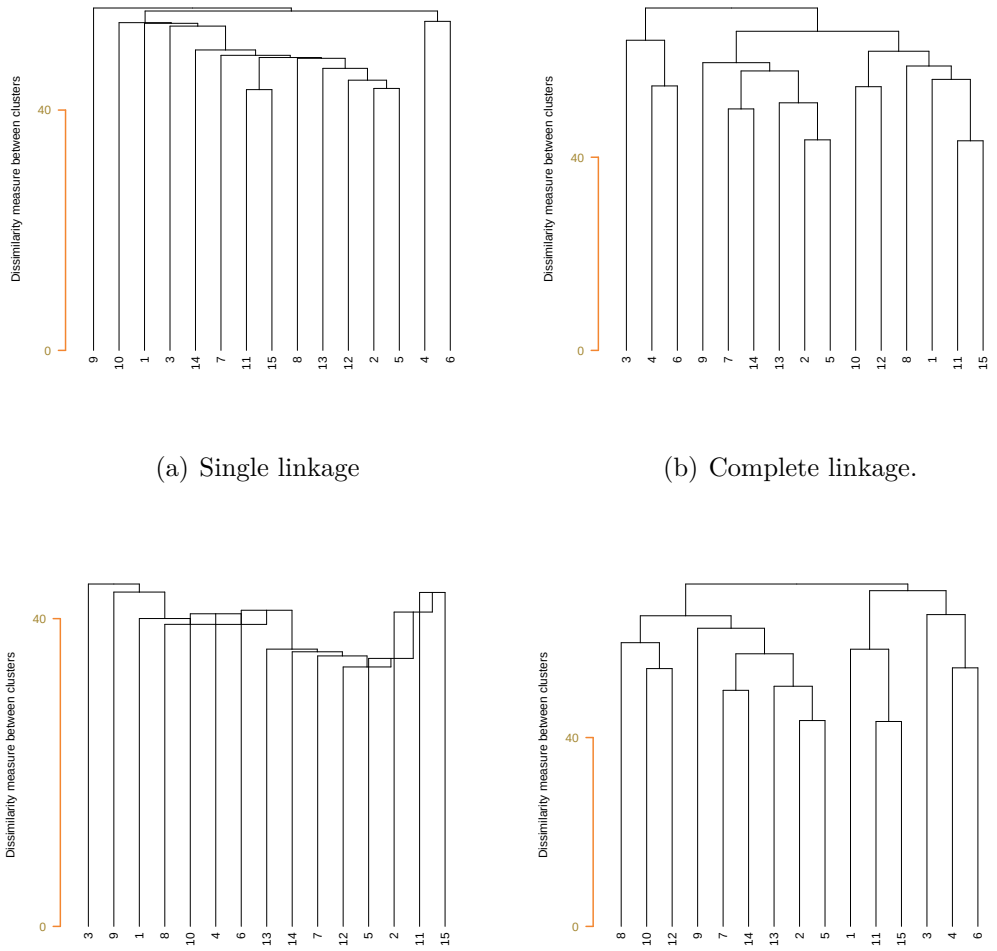
2.1.6 Median Linkage

Median linkage, also called the weighted pair-group centroid method, is a variation on centroid linkage which defines the distance between two clusters as the weighted distance between their centroids. This weight is corresponding in size to the number of samples in each cluster. This method is only used with Euclidean distance. This linkage can be used for downweighting the effect of outliers by using the median instead of the mean see Figure

2.3(c). In the median linkage, the distance between two disjoint clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as,

$$D_{\text{median}}(\mathcal{C}_1, \mathcal{C}_2) = ||\tilde{\mathbf{y}}(\mathcal{C}_1) - \tilde{\mathbf{y}}(\mathcal{C}_2)||^2$$

where $\tilde{\mathbf{y}}(\mathcal{C}_1)$ and $\tilde{\mathbf{y}}(\mathcal{C}_2)$ are the medians of clusters $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.



(a) Single linkage

(b) Complete linkage.

(c) Median linkage.

(d) Ward's linkage.

Figure 2.3 Dendrograms corresponding to the four different linkages in hierarchical clustering applied to random data. As it is shown in Figure 2.3(c) monotonicity property is not satisfied for all linkages.

2.1.7 Properties of hierarchical algorithms

Properties of hierarchical algorithms are usually expressed as (i) Lance-Williams, (ii) Monotonicity, and (iii) Space Distortion (conserving, contraction, or dilating) (Rencher, 1998). Lance-Williams property is discussed with more details in Chapter 4 with its extension for forestogram framework.

Monotonicity property of clustering states that a cluster is not allowed to get merged with another cluster at a height that is less than the height of the previously combined clusters. This also is referred to as *ultrametric* which inspires the separability assumption for our developed forestogram in the context of extended hierarchical algorithms for biclustering. In Figure 2.3(c) median linkage as an example of nonmonotonic is shown, similarly with a counter example one can demonstrate centroid is not monotonic either.

Properties of the space of distances can change after creation of clusters. Clustering algorithm is space-conserving if the spatial properties always stay intact, otherwise the space can be either contract or dilate in the sense of changes occur to the distances between any arbitrary pair of data points. The tendency of a singleton clusters to join the newly created cluster is called contraction, while the opposite behavior is known as dilating. In space contracting algorithms, larger clusters frequently appear after each merge, so that singletons eventually combine with non-singleton large clusters. For the space-dilating algorithms we expect to have more new clusters rather chaining property (Rencher, 1998). In this fashion, singleton clusters are more likely to join the other singleton clusters rather than with non-singleton ones. Let's consider three clusters, $\mathcal{C}_i$, $i \in \{1, 2, 3\}$, where the pairwise distances are defined as,

$$D(\mathcal{C}_1, \mathcal{C}_2) < D(\mathcal{C}_1, \mathcal{C}_3) < D(\mathcal{C}_2, \mathcal{C}_3) \tag{2.1}$$

If the equation in (2.1) is not held, the clustering algorithm is space-contracting. Space-conserving algorithm does meet the conditions in equation (2.2).

$$D(\mathcal{C}_1, \mathcal{C}_3) < D(\mathcal{C}_{\{1,2\}}, C_3) < D(\mathcal{C}_2, \mathcal{C}_3) \tag{2.2}$$

And space-dilating algorithm does not satisfy the equation (2.2).

The hierarchical algorithm with single linkage is prone to space-contracting tendency because of violating the first inequality $D(\mathcal{C}_{\{1,2\}}, C_3) = \min\{D(\mathcal{C}_1, \mathcal{C}_3), D(\mathcal{C}_2, \mathcal{C}_3)\} = D(\mathcal{C}_1, \mathcal{C}_3)$, therefore single linkage is not preferred in many fields see Figure 2.3(a). On the other hand, complete linkage is in the class of space-dilating algorithms because of violation of the second inequality $D(\mathcal{C}_{\{1,2\}}, C_3) = \max\{D(\mathcal{C}_1, \mathcal{C}_3), D(\mathcal{C}_2, \mathcal{C}_3)\} = D(\mathcal{C}_2, \mathcal{C}_3)$ correspondingly new clusters are more likely to be seen at each iteration see Figure 2.3(b). Other hierarchical linkages are often somewhere in between single linkage and complete linkage, e.g. centroid linkage and

average linkage algorithms are more biased to space-conserving, however Ward's linkage is in favor of space-contracting see Figure 2.3(d) (Rencher, 1998). The mentioned space properties provide a clue for considering how likely a dendrogram one can expect in terms of balanced or unbalanced group of homogeneous data.

Furthermore, stability and convergence of hierarchical clustering algorithms are discussed in Carlsson and Mémoli (2010). Hartigan consistency is reviewed in Eldridge *et al.* (2015) as a framework for analysing the hierarchical clustering. Then merge distortion metric is suggested in order to alleviate the over-segmentation and improper nesting with two limited properties, separation and minimality (Eldridge *et al.*, 2015).

2.1.8 Model-based cluster estimation

The majority of clustering algorithms can be divided into distance-based methods or model-based methods. Distance-based techniques are easy to understand and simple to implement. On the contrary, model-based approaches are flexible and adapt to complex data patterns, but are counter intuitive to implement. In model-based clustering a family of statistical models is considered for data. Estimating the number of clusters in both approaches is a complex problem. However, some methods are developed for distance-based methods using cross validation (Tibshirani *et al.*, 2001), or often asymptotic model selection criteria is used (Claeskens and Hjort, 2008). Estimating the number of clusters through cutting the dendrogram at certain height, is equivalent to find a tangible gap on the height of the dendrogram for a natural grouping. An approximate model selection criteria such as AIC (Akaike, 1973) or BIC (Schwarz, 1978) can be applied to cut the dendrogram if a statistical model is used to produce the nested clusters (Heller and Ghahramani, 2005; Heard *et al.*, 2006). We further extend the idea of model-based cluster estimation in Chapter 4, for finding the number of biclusters on our suggested forestogram. We suggest a model selection method which finds the cutting point of the forestogram automatically. In order to achieve this goal, suppose that $\boldsymbol{\theta}$ is the associated parameter of the clusters. It turns out that $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood of the data if $\boldsymbol{\theta}$ was known exactly. As far as, $\boldsymbol{\theta}$ has not yet known, and grouping of the data is our main concern rather than the value of $\boldsymbol{\theta}$, the predictive distribution of the data can contribute to find the optimal cluster assignments. The solution of this predictive distribution is computed by marginalizing the likelihood of the data multiplied by prior distribution of $\boldsymbol{\theta}$ over the clustering parameter, i.e. $p(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$. This resembles a BIC criterion whose optimal number of clusters is found by computing it over all levels of the tree on a given forestogram. The marginal provides a measure of a merge that is supported by the model if increases for that merge. This way forestogram can decide how to cut the forestogram.

2.1.9 Biclustering

One of the desired goals in data analysis for an arbitrary multivariate dataset is to find barycentric relations for the hidden structure among subjects and their corresponding attributes (Govaert and Nadif, 2013). Finding the partitions of rows and columns at the same time is known as biclustering, however, in the literature is also referred to as two-mode clustering, coclustering, simultaneous clustering, two-way clustering, or two-side clustering, block clustering (Govaert and Nadif, 2013). Since the late 90s, biclustering has been the term most widely used in bioinformatics (Govaert and Nadif, 2013). The simple and easiest way for biclustering is to perform a clustering algorithm to both sides, rows and columns independently to find the relevant blocks. Since for multivariate data analysis, columns of the matrix are generated from the samples located on the rows, independent clustering of rows and columns is not statistically significant for real world problems. After unfolding the relations among the correlated rows and columns, visualizing the simultaneous partitions is the next issue that should be addressed for biclustering problem. *Heatmap* is a conventional visualization method to display a matrix in terms of biclustering. Although heatmap applies hierarchical clustering on rows and columns independently, but it provides a good visualization scheme that is easy to understand. Independent dendrograms on row and column demonstrate a biclustering visualization where the intersection of row clusters in conjunction with that of column illustrates the structure of blocks underlying the data.

Despite the complicated nature of biclustering, this viewpoint to unsupervised partitioning has been arisen in many applied fields such as topic modeling in natural language processing whose goal is to extract the topics from the corpus of documents (Rugeles *et al.*, 2017; Orzechowski and Boryczko, 2016), web mining to reveal the web pages that are viewed by certain group of people (Rathipriya and Thangavel, 2014), recommender system and marketing as a general class of problems where a shared behavior among group of people is the case of interest (Wang *et al.*, 2015; Alqadah *et al.*, 2015), bioinformatics to gene expression profiling for molecular, cell or tissue in biological measurements (Eisen *et al.*, 1998; Eren *et al.*, 2013), manufacturing systems to show how the processing time and available machines as resources are interacting together (Boutsinas, 2013; Liiv, 2010), public transport for evaluating the most important roads in the network (Freiria *et al.*, 2015; Owens, 2009) etc. Toward the meaningful biclustering viewpoint that aims at finding a block of correlated rows and columns, pairwise distance known as dissimilarity matrix plays a central role in classical approaches. The modern methods can be divided into three categories: (i) Bayesian approach: this perspective assumes a prior distribution over the statistical parameters of model (Gu and Liu, 2008; Martella *et al.*, 2008; Zhang, 2010) ; (ii) Frequentist aspect: in this view the statistical model underlies fixed unknown parameters, such as mixture models (Lazzeroni and

Owen, 2002) ; and (iii) Matrix approximation view: this approach reconstructs the original matrix by multiplying two low-rank matrices where the first multiplicand and the second one reflect the cluster assignment of subjects and attributes, respectively (Donoho and Stodden, 2004; Ding *et al.*, 2005; Arora *et al.*, 2012; Wang and Zhang, 2013; Gillis, 2011; Klingenberg *et al.*, 2009; Cai *et al.*, 2008; Lee and Seung, 2000).

Most of the biclustering algorithms are developed based on fixed number of biclusters that ask the user to manually gives this information to the algorithm. In many applications, determining the number of biclusters is another issue that needs to be handled by the algorithm itself. The pattern that biclustering algorithm is seeking to find falls into three main categories namely, constant, additive and multiplicative. The pattern of equal values constitutes the constant model. If the rows and columns share an additive factor then the pattern is denoted by additive term, while the multiplicative model requires multiplicative factors to represent the bicluster pattern. Similarly, it is not hard to imagine a pattern that integrates these three different models to define the bicluster notion. From the statistical point of view, a pattern consisting of the correlation between rows and columns is preferred (Madeira and Oliveira, 2004). Since there is no concrete choice for defining a bicluster, the criterion for identifying a submatrix as a bicluster is problem dependent. In this regard, there is no algorithm that is able to detect all variations, thus for a particular problem or certain type of data, we need to design a new procedure to account for the required specifications. To this end, we design the forestogram framework for computing and visualizing the hierarchical biclustering introduced in Chapter 4 with successful results in two major fields (Ghaemi *et al.*, 2017c). In the following, we go through the definition, specification, and goals of biclustering and related research works in public transport and bioinformatics.

2.2 Application

In a number of complex systems such as public transit network, biological sequencing and in general time series observations, similar groups are expressed in a nested hierarchical structure (Tumminello *et al.*, 2010; Aghabozorgi *et al.*, 2015). According to the intrinsic seasonal trend that repeats itself systematically over time, subclusters of homogeneous entities can be found in the underlying data up to a certain level. Moreover, often, in real world time series clustering analysis, determining the number of exact similar groups is a tough issue. For this reason, hierarchical approaches are one of the common choices for time series clustering in addition to the strength of visualization power in terms of binary dendrogram tree (Aghabozorgi *et al.*, 2015; Van Wijk and Van Selow, 1999; chung Fu, 2011).

2.2.1 Public Transport

The importance of the public transportation and its influence in the real life of many people in large cities around the world, rises a new family of problems that is not confined into a particular branch of science. Hence, usage of the smart card data creates the opportunity for several different researchers from diverse disciplines e.g. data mining, machine learning, urban computing and planning, management, business, civil engineering, industrial engineering, statistics, mathematical engineering, geographic information system (GIS), etc. to outreach and extend their methods to analyze the data for the public transport authorities. Figure 5.14 shows a typical public transit network including users, buses, and subway lines.

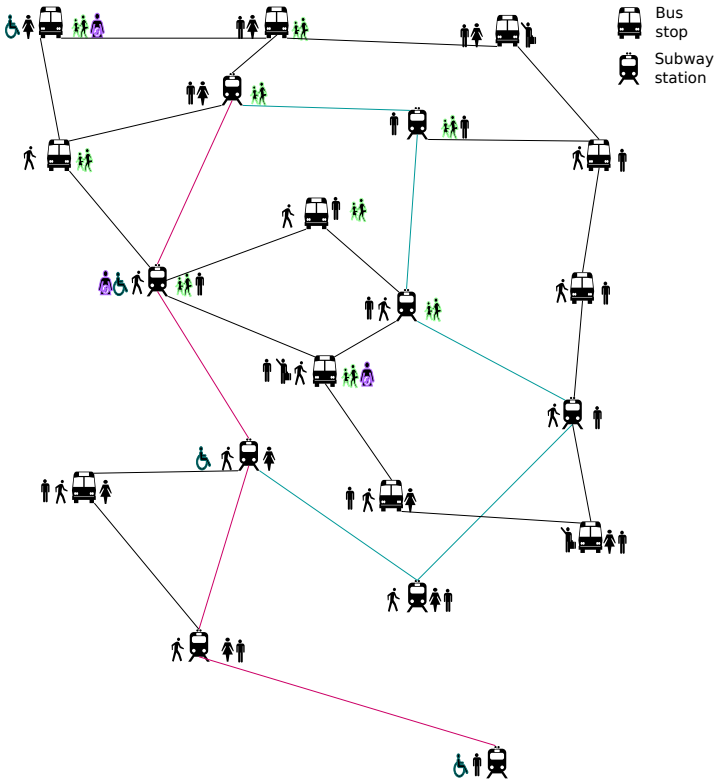


Figure 2.4 A typical public transit network.

Despite extensive researches have been done on public transportation domain, various obstacles have been arisen for specific purposes which require particular approaches to address them. Here we review a recent concerning problem of clustering the transit users according to the spatial-temporal data gathered from smart cards to analyze their behavior in the public transit network.

Smart card data, contains worthwhile digital information of daily locations visited at certain period of a large number of individuals (Pelletier *et al.*, 2011). Beside other sources of

information such as mobile phone, GPS tracker vehicle, e.g. bike, car, motorcycle, credit card transactions, social network, and many other sources of information gathering, smart card data is a promising source of users digital information. Thus, this helpful information could be utilized to characterize and model urban mobility patterns (Hasan *et al.*, 2012). Other useful information such as travel time and number of passengers for the sake of congestion analysis and planning improvement, could be possibly extracted as well (Fuse *et al.*, 2012).

Smart card data, usually provides two distinct information; spatial and temporal (Pelletier *et al.*, 2011). Spatial data consists of coordinates of the bus stop e.g. latitude and longitude that could be GPS data or relative values. Temporal data describes the time each trip is taken, this information could be encoded in a $0-1$ vector, where start of the trip is indicated by 1. According to these information, analyzing users behavior is divided into three categories, 1) Spatial patterns, 2) Temporal patterns and 3) Spatial-Temporal patterns.

1. In the first case, methods of analyzing spatial pattern, are taking the bus/subway stop's information into account. It turns out measure of behavioral pattern only depends on the location of stops, taken by the users rather than having known the starting hour of their trip.
2. The second methods seek the information pertinent to the temporal data associated to the public transport usage. Consequently, computing user similarity score is carried out regardless of geographical information. The indices of 1 occurrences in the encoded vector, are playing the central role in this approach.
3. The third scenario, is a mixture of the spatial and the temporal data, called spatial-temporal data analysis to investigate users' behavior. It could be viewed as a combination of the last two steps or an independent approach to recognize the spatial-temporal behavioral pattern in the public transport domain.

Hierarchical algorithms have been used as an unprecedented clustering method on spatial-temporal public transit data, such as shared bicycle policy analysis (Lathia *et al.*, 2012), traffic mining (Froehlich and Krumm, 2008), spatial-temporal clustering for congestion patterns detection in urban road network (Anbaroglu *et al.*, 2014), rush hour motorcycle flow data analysis in Taipei City, Taiwan (Feng *et al.*, 2014). Shirui (2016) uses hierarchical clustering for visual analysis of the spatial-temporal traffic flow patterns generated from transport hubs in Shanghai. Divisive analysis clustering is suggested for classification of large amount of speed data collected from GPS receiver in India (Patnaik *et al.*, 2016). Hierarchical methods show successful result in extracting the temporal patterns from the trip data through the Beijing subway system to characterize individual passenger movement patterns (Xu *et al.*, 2016).

2.2.2 Bioinformatics

Nowadays, according to the emerge of promising technologies such as genomic and proteomic, and metagenomic developing modern tools and powerful algorithms is extremely necessary (Clarke *et al.*, 2008; Huber *et al.*, 2015; Tyanova *et al.*, 2016). Therefore, extracting the knowledge from these multiomics biological dataset, can help improving the level of human health (Conesa *et al.*, 2016; Norris *et al.*, 2017; Ritchie *et al.*, 2015). There are three main approaches to address this issue in order to analyze the biologically relevant knowledge, supervised, unsupervised and semi-supervised (Serra *et al.*, 2015; Maetschke *et al.*, 2013). For supervised viewpoint, we have to spend fairly huge amount of financial budget to associate a label to each sample with respect to certain disease or health related problems (Zhao *et al.*, 2008). However, from the unsupervised perspective regardless of the case study, we can provide a general knowledge about the interaction of certain biological measurement and patients for studying the racial, epidemiological, environmental, sociodemographic, etc. factors that may have causal effect or correlation with a number of common diseases or medical conditions. Moreover, unsupervised preprocessing the data can affect the future supervised analysis by adjusting the noise level or removing the outliers (Huang *et al.*, 2017). Additionally, semi-supervised methods that are considered as a combination of formerly introduced approaches can be used effectively in the case of scarcity of the labels (Hassanzadeh *et al.*, 2016). In this work, we aim to analyze the pregnancy across diverse multiomics profiling technologies according to supervised information expressed as gestational age and unsupervised way to investigate the biological relations between the multiomic measurements and patients that are sampled in three different trimesters of pregnancy. In this regard we choose the generalized linear model via penalized maximum likelihood as a method of supervised variable selection to predict the gestational age. Furthermore, we investigate the performance of forestogram for biclustering task where unsupervised selected variables correlated with patients can show how much supervised information contributes to the prediction of pregnancy trimesters.

CHAPTER 3 RESEARCH APPROACH AND STRATEGY

In this chapter, we review the main goals of the forestogram framework for hierarchical biclustering to outline how the suggested methodological development can be investigated through the applied projects as the contributions of this research. Then this research methodology and its application are justified by presenting a big picture of two applied projects in public transit and bioinformatics in addition to a general purpose methodology with published and submitted journal articles.

Finding a tractable solution for NP-hard problems with reasonable approximation is a fundamental question in theoretical development of a new methodology. While collecting supervised information for a dataset potentially incurs fairly huge overload of budget to any research project, devising new techniques for analyzing the data becomes viable through unsupervised learning. Grouping similar patterns for a given dataset without human knowledge is considered as a big advantage toward smart data analysis. In this regard, clustering is the first choice that exists for knowledge discovery from the data without label information. However, clustering is an optimization problem which falls into the NP-hard class of algorithms that requires a good approximation to solve the problem efficiently. Among the two approaches for clustering, we propose the hierarchical methods as an informative outlook for experts to consider for unsupervised grouping of data that provides a general perspective of formation of the clusters incrementally. Despite appealing visualization guide for clustering that hierarchical algorithms deliver to the users, determining the number of clusters according to this abstraction of groups is not trivial. Moreover, for many applied projects only grouping of samples is not enough. For high dimensional datasets, grouping of the attributes is another important question that is already addressed in terms of supervised learning such as feature extraction, sparsifications, dimension reduction and variable selection. These questions, e.g., grouping of observations, selecting the variables, number of groupings and visualization of the result can be addressed separately in the context of independent projects. Since, each question eventually has an estimated solution, the overall solution as a sequence of diverse algorithms is highly prone to diverge from the optimal solution. Therefore, designing a unified framework based on concrete foundation reduces the risk of approximation error while makes it easy to analyze the algorithm and the mathematical properties to describe the problem.

Research Questions

The main objective of this research is to develop a general purpose methodology for biclustering with hierarchical aspects. In Figure 3.1, a big picture is given to show how we define two applications with relevant links to our developed forestogram framework for

biclustering. In the following, we answer a number of research questions from methodological and applied viewpoints,

RQ1: Why do we suggest hierarchical clustering approach for the applied projects ?

In practical applications, a method that is easy to understand by people of the field that are already familiar with is hierarchical clustering. For many people in industrial engineering and bioinformatics, hierarchical methods are easy to understand because of the greedy agglomerative nature of the algorithm and its visualization power. This way, expert people can feel more comfortable to use the algorithm because they understand the mechanism which produces the result in the background. Additionally, the binary tree known as dendrogram elaborates the result such that interpreting the result is easy enough to provide a descriptive summary of the data.

RQ2: How can we figure out the number of clusters for a given data ?

This is a major question in hierarchical clustering. Basically, dendrogram demonstrates a good overview of the shape and relation of the clusters through a hierarchical abstraction where the height of each merge denotes the dissimilarity of the newly formed clusters. However, for complex data with high dimension this is not the best solution. Moreover, for some scenarios there would be more than one apparent cut on the tree so that determination of the estimated height to cut the tree is necessary. We introduce an information criterion measure to determine the significant number of groups in the hierarchical setting called FORIC to help data analysts finding a relevant point for cutting the dendrogram or forestogram.

RQ3: How can we illustrate the result of clustering ?

Dendrogram is a conventional mode to expose the hierarchical structure of the clusters in terms of binary rooted tree. However, dendrogram is capable of presenting the result for grouping of samples. If one is interested to explore the relation of samples and attributes simultaneously, we suggest to add an extra orthogonal merge to the tree to represent the merge between two different groupings coming from observations and corresponding features. The advantage of this visualization is to conform the notion of dendrogram in the augmented space.

RQ4: What is the best way to aggregate both clustering and variable selection ?

Traditionally, after variables are selected by preprocessing the data, clustering algorithms are performed on the altered data. Theoretically, analysis of this two step procedure is not well understood since variable selection and clustering methods are not coherently developed from the same principals. We suggest to group samples and features with the same algorithm so that comprehension of the algorithm is still easy to understand yet provides a unified

framework to study the properties of the method for explaining the extracted biclusters.

RQ5: Why is model based clustering the best approach for biclustering ?

We are seeking a unified framework to perform the biclustering so that inference could be done meaningfully. We prefer model-based clustering so that the uncertainty of the method is quantifiable. Moreover, a model provides a link to connect sample clustering to feature clustering. In addition, the model has enough flexibility to answer more questions such as the number of clusters, importance of clusters and modularity of the data by simple inference on the model. For instance, to determine the number of clusters, we marginalize the clustering parameter defined in the model to derive the predictive distribution. Therefore, we can identify the cutting point on the dendrogram or forestogram by computing the optimal posterior empirically through the data. This way, the model allows the data to lead the algorithm to answer any question pertinent to the clustering problem.

RQ6: How can we use hierarchical clustering in the public transit domain ?

For hierarchical algorithms, we need to define a dissimilarity measure as the input parameter. However, in public transit data, we often deal with two types of distinct information, temporal and spatial. For the temporal data that are similar to time series data, the normal hierarchical clustering techniques are not suitable because off-the-shelf distance metrics are not designed for binary vectors. To this end, we first suggest a projection technique to map a long binary vector of temporal usage into three dimensional space which retains the proximity of pairwise similarity. Additionally, this projection helps better understanding the temporal patterns visually on a semi-circle trajectory. For the next step, we take the temporal information as a latent variable to extract the spatial-temporal patterns from the data such that Euclidean metric becomes feasible because of geodesic property of GPS location history. By the temporal projection and modification of spatial-temporal distance, we can use the designed forestogram to find the similar group of public transit subscribers with the underlying spatial and temporal patterns. It is worth to mention that, with the suggested FORIC, we can also identify the number of similar groups of people in the transit network as well as the number of existing patterns.

RQ7: How can we integrate multiomics datasets for pregnancy ?

First of all, we address the issue of pregnancy in terms of seven independent multiomics dataset. We analyze each dataset separately to show how each biological measurement is influencing the term of pregnancy. For the integrative model, we suggest the stacked generalization and forestogram approaches for supervised and unsupervised integrative data analysis. According to our experiments, the first two trimesters of pregnancy need supervision information for accurate prediction by a classifier algorithm, while forestogram is empowered to uncover the third trimester in pregnancy. The main advantage of forestogram is that

the visualization comes with the method which provides a profound insight for biologists to investigate the interaction of different omics with three terms of pregnancy. We perform a multiomics analysis of 51 samples from 17 pregnant women, delivering at term. The datasets include measurements from the immunome, transcriptome, microbiome, proteome, and metabolome of samples obtained simultaneously from the same patients. Pregnant women presenting to the obstetrics clinics of the Lucile Packard Children's Hospital at Stanford University for prenatal care were invited to participate in a cohort study to prospectively examine environmental and biological factors associated with normal and pathological pregnancies. Women were eligible if they were at least 18 years of age and in their first trimester of singleton pregnancy. Samples were obtained during the first ($7--14$ weeks), second ($15--20$ weeks), and third ($24--32$ weeks) trimesters of pregnancy, and 6 weeks post-partum.

The contribution of this thesis is mainly based on the hierarchical biclustering framework with forestogram visualization as it is shown in Figure 3.1. In this regard, we introduce the problem of hierarchical biclustering in Chapter 4, by elucidating how hierarchical model generation is connected to Bayesian viewpoint. This new approach helps us applying the statistical inference to study the properties of the model, such as model selection with FORIC to provide a clue for determining the number of biclusters in the data. Then we show under what conditions on the data the existing separable biclusters can be found effectively through the model. Then with Lance-William trick it is easy to reduce the computational time complexity of the algorithm so that one can simply run this algorithm on a personal computer for a fairly medium size data matrix with our efficient implementation. Eventually, with forestogram 3D visualization, this algorithm demonstrates the interrelations among the rows and columns of the data matrix expressed in the form of biclusters intuitively with more explanation for the end users. The first manuscript is submitted in Ghaemi *et al.* (2017a), and additionally the beta version of `R package` is available in Ghaemi *et al.* (2017c). In addition to the simulation study and yeast galactose example in Chapter 4, two other applications from different field of applied science are considered in this thesis to investigate the performance of the suggested data analysis toolbox for unsupervised data exploration. Public transportation data is the first inspiration of this algorithm to use the hierarchical biclustering for revealing the hidden spatial-temporal pattern underlying the smart card data. The temporal analysis is carried out by a novel semi-circle projection to map the high-dimensional data onto the 3D space of time with clock-like behavior for the temporal part of the data. This work is published in Ghaemi *et al.* (2017b). In the next step, hourly usage acts as a latent variable to link the conventional Euclidean measure of distance to the geodesic spatial information in order to use the forestogram for analyzing the spatial-temporal properties of smart card data in Chapter 5. The second application deals with the gestational age prediction in the

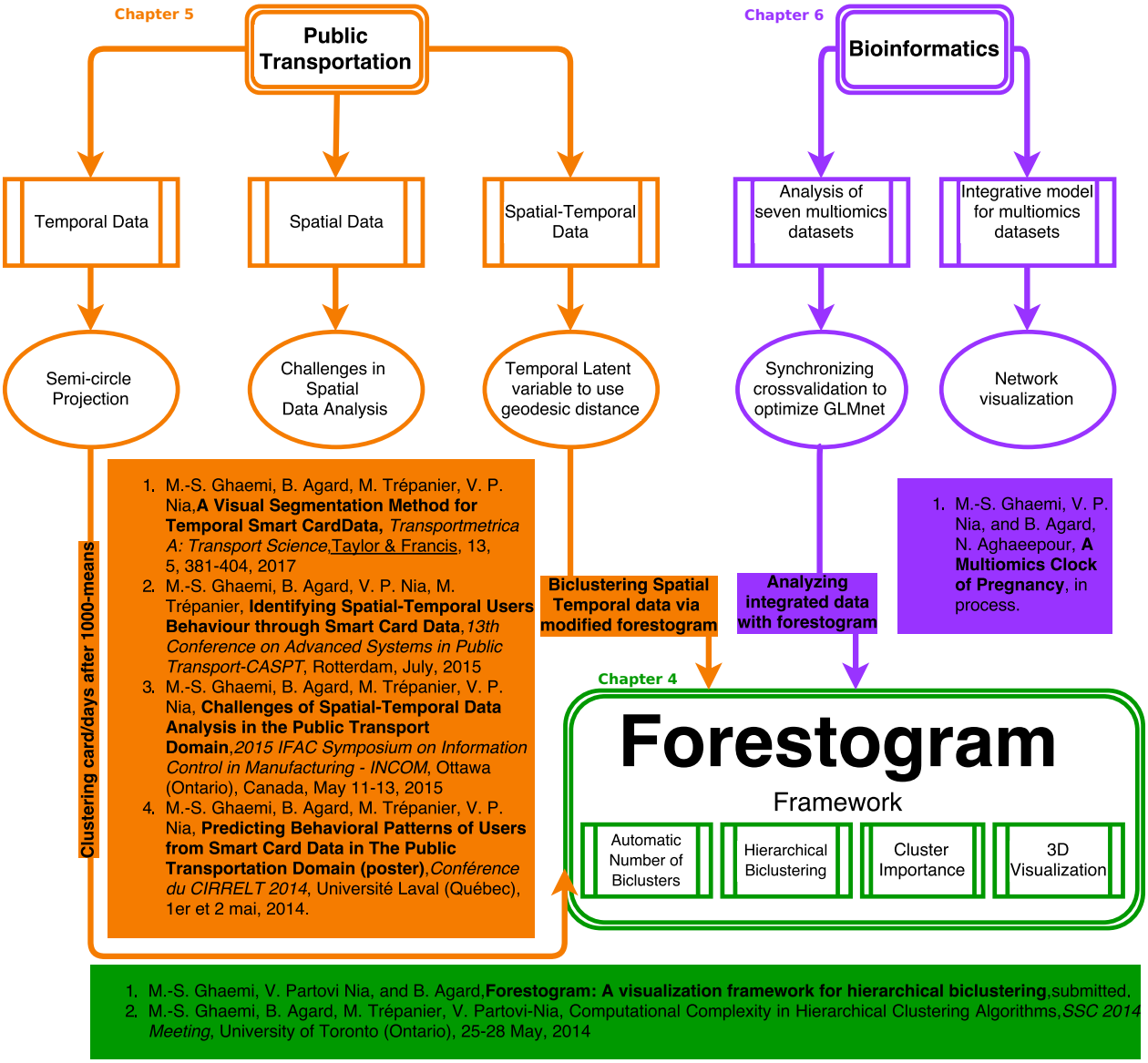


Figure 3.1 Thesis contribution.

bioinformatics area. Seven different multiomic datasets constitute the input data to predict the gestational weeks. Typically, this integrative data analysis requires supervised response variable to make the prediction with. However, selecting the features across seven datasets is also needed to describe the model through biological pathways for clinical studies. We use the forestogram along with elastic net to peruse an integrative biological model for predicting the gestational age and presenting the influential features that affect the patients during the term of pregnancy. This work is available in Chapter 6.

CHAPTER 4 ARTICLE 1: FORESTOGRAM: A VISUALIZATION FRAMEWORK FOR HIERARCHICAL BICLUSTERING

4.1 Abstract

Many biological datasets such as microarrays, metabolomics, and proteomics involve observations (or subjects) in rows, and attributes (or genes, metabolites, proteins) in columns. Often simultaneous grouping of rows and columns, i.e. biclustering, is desired. Each bicluster consists of a group of observations highly correlated in a group of attributes. Despite great efforts on developing biclustering algorithms, a proper visualization seems to be lacking in the literature. A visualization tool helps practitioners understanding how biclusters evolve. Here we provide this tool using *forestogram*. Forestogram combines rows or columns iteratively towards constructing a forest over a collection of dendrograms with a common root. We develop a simple strategy for extracting natural biclusters by cutting the forest using a simple information criterion. The effectiveness of our technique is tested on simulated data, and on real data. [1]

Keywords Biclustering, dendrogram, hierarchical clustering, linkage

4.2 Introduction

Clustering, or data grouping, is a challenging problem. Clustering is NP-hard, i.e. the number of different ways to group data grows exponentially with the sample size. Clustering algorithms can be categorized into two categories: hierarchical, and partitional. Hierarchical methods find the nested clusters recursively, while partitional approaches provide only a single grouping. Partitional algorithms require the number of clusters to be set a priori. Hierarchical approaches, on the contrary, starts from each item as a singleton and builds clusters until all data fall in a single cluster. A clustering algorithm that assumes a statistical model for clustering data, is called model-based clustering (McLachlan *et al.*, 2004). Practitioners often prefer hierarchical clustering, because of the visual guide produced through dendrogram. Clustering *linkage*, also known as *dissimilarity*, plays a central role in building the dendrogram.

Biclustering, also known as *coclustering*, and *joint clustering* is a general class of methods that aims to partition a data matrix. Unlike clustering that groups observations, *or* attributes, biclustering searches a grouping on observations *and* attributes at the same time. The

1. MS. Ghaemi, B. Agard, V. Partovi Nia. "Forestogram: A visualization framework for hierarchical biclustering", *Statistical Analysis and Data Mining*, submitted.

advent of high-dimensional data calls for devising new algorithms to exploit the clusters more effectively.

Biclustering attracted researchers from various fields because of its modern applications (Zhang, 2010). Biclustering is used to cluster documents and words in text mining (Orzechowski and Boryczko, 2016), genes and experimental conditions in bioinformatics (Eren *et al.*, 2013), tokens and contexts in natural language processing (Tu and Honavar, 2008), users and movies in recommender systems (Xu *et al.*, 2012), etc. The first joint clustering method appeared in statistics literature in Hartigan (1972), but implemented after few decades (Cheng and Church, 2000). Like clustering, biclustering involves two cultures i) statistical approach that assumes a probabilitic distribution (Sheng *et al.*, 2003; Gan *et al.*, 2008; van Uitert *et al.*, 2008; Lazzeroni and Owen, 2002; Sheng *et al.*, 2003; Gan *et al.*, 2008) ; and (ii) the algorithmic approach that minimizes a dissimilarity (Hartigan, 1972; Hochreiter *et al.*, 2010; Martella *et al.*, 2008), for a comprehensive review see Busygin *et al.* (2008).

Most of the biclustering techniques are partitional and the number of blocks is the input of the algorithm. However, in a number of applications hierarchical approach is very common, because of two main advantages: i) having little assumption on data and number of groups ii) providing a visualization diagram through the dendrogram.

A simple hierarchical biclustering method is known as *heatmap*, which produces two independent dendrograms, one on rows and another on columns. This representation is loose due to the independent construction of row and column groupings. However, an interesting visualization tool for biclustering is proposed using convex reformulation of the biclustering problem in Chen *et al.* (2015), but it lacks the conventional dendrogram representation that practitioners are used to see. An agglomerative method using a complex Bayesian model is suggested in Fowler and Heard (2012). Smith *et al.* (2008) argues that complex models may lead to junk clusters if agglomerative method is used.

We propose i) a natural extension of biclustering method using common linkages, ii) produce forestogram, a conventional graphical tool that extends dendrogram, iii) benefit a simple hierarchical model to develop a criterion as a reference for cutting forestogram. It turns out that our criterion is the natural biclustering extension of the well-known information criteria, such as the AIC (Akaike, 1973) and BIC (Schwarz, 1978).

The paper is structured as follows. Section 4.3 describes our proposed methodology and forestogram. Section 4.4 studies the computational complexity of the forestogram construction. Section 4.5 compares forestogram with some common biclustering methods, and Section 4.6 shows the application of forestogram on the yeast galactose data.

4.3 Hierarchical Biclustering

Hierarchical biclustering is a natural extension of hierarchical clustering for grid matrices. Section 4.3.1 generalizes common linkages for biclustering. Section 4.3.2 explains how to build the forestogram using the generalized linkage. Section 4.3.3 develops an information criterion to provide a statistically meaningful suggestion for the forestogram cut, and Section 4.3.4 explores the relationship between biclustering and forestogram.

4.3.1 Bilinkage

Hierarchical biclustering algorithms require a dissimilarity measure to merge block of clusters and build nested groups. The dissimilarity measure is a positive semi-definite symmetric mapping of pair of groups, onto real numbers. Dissimilarity, however, may not satisfy the triangle inequality unlike the distance. The common linkages include single linkage or nearest neighbors, complete linkage or farthest neighbors, average linkage, centroid linkage, median linkage, and Ward's linkage, see (Sørensen, 1948; Sokal, 1958; Eisen *et al.*, 1998; Murtagh and Legendre, 2014) for more details.

The linkage is defined using a distance, often the Euclidean distance, but may be defined on metrics such as Manhattan, Chebyshev, or Mahalanobis distance.

We suppose grid biclusters, and use I to index row clusters, and J to index column clusters. The first step in building the hierarchical biclustering is to generalize the linkage to a *bilinkage* to measure the dissimilarity between matrix blocks. Any merge, however, cannot be visualized by a nested tree. Therefore, a convenient bilinkage must be defined over a pair of biclusters, using row and column directions. Suppose $\mathcal{C}_1$ and $\mathcal{C}_2$ are disjoint rectangular biclusters,

$$\text{Bilinkage}(\mathcal{C}_1, \mathcal{C}_2) = \min_{I \neq I', J \neq J'} \left\{ D(\mathcal{C}^{\text{row}}_{I1}, \mathcal{C}^{\text{row}}_{I'2}), D(\mathcal{C}^{\text{col}}_{1J}, \mathcal{C}^{\text{col}}_{2J'}) \right\} \tag{4.1}$$

where $\mathcal{C}^{\text{row}}_{I1}$ is the Ith row-cluster of bicluster $\mathcal{C}_1$, $\mathcal{C}^{\text{col}}_{1J}$ is the Jth column-cluster of bicluster $\mathcal{C}_1$, and D is a clustering linkage. Table 4.1 gives the definition of the commonly used linkages. The minimum in (4.1) is taken once over a pair of row-clusters, and once over a pair of column-clusters. This minimum defines the direction of the merge, a row merge, or a column merge. We suppose that data are standardized, so that row and column blocks are comparable.

4.3.2 Forestogram

Forestogram is a collection of binary trees that consists of multiple hierarchical dendrograms. Construction of the forestogram is bottom-up, such that a pair of row-wise or column-wise clusters is combined together at each level by starting from singleton clusters.

Table 4.1 A list of common linkages for hierarchical clustering, defined using the Euclidean distance, where $\bar{\mathbf{y}}$ denotes the mean, and $\tilde{\mathbf{y}}$ denotes the median.

Linkage	Definition
Single	$\min_{\mathbf{y}_i \in \mathcal{C}_1, \mathbf{y}_j \in \mathcal{C}_2} \|\mathbf{y}_i - \mathbf{y}_j\|$
Complete	$\max_{\mathbf{y}_i \in \mathcal{C}_1, \mathbf{y}_j \in \mathcal{C}_2} \|\mathbf{y}_i - \mathbf{y}_j\|$
Average	$\frac{1}{\|\mathcal{C}_1\|\|\mathcal{C}_2\|} \sum_{\mathbf{y}_i \in \mathcal{C}_1} \sum_{\mathbf{y}_j \in \mathcal{C}_2} \|\mathbf{y}_i - \mathbf{y}_j\|$
Ward	$\frac{\|\mathcal{C}_1\|\|\mathcal{C}_2\|}{\|\mathcal{C}_1\|+\|\mathcal{C}_2\|} \|\bar{\mathbf{y}}(\mathcal{C}_1) - \bar{\mathbf{y}}(\mathcal{C}_2)\|$
Centroid	$\|\bar{\mathbf{y}}(\mathcal{C}_1) - \bar{\mathbf{y}}(\mathcal{C}_2)\|$
Median	$\|\tilde{\mathbf{y}}(\mathcal{C}_1) - \tilde{\mathbf{y}}(\mathcal{C}_2)\|$

Forestogram merges a block of rows or a block of columns in each step, depending on the direction that minimizes the bilinkage (4.1). After each merge, the dissimilarity measure is recomputed to identify the next merge direction. This approach, gives a new block of data on the forestogram. A grouping is extracted if the forestogram is cut at a certain height, see Figure 4.1.

Forestogram has a number of interesting advantages to interpret the block-clusters of data as follows. Each bicluster reflects the order of rows and columns that shares a similar pattern. The merge path gives a visual representation on the evolution of the biclusters. Forestogram gives a visual guide on the interaction between row and column groupings. A row dendrogram and a column dendrogram can be extracted by projecting the forest over rows and columns, see Figure 4.2. The last property is attractive for practitioners who are familiar with heatmap graphics.

4.3.3 Number of Biclusters

Estimating the number of biclusters through cutting the forestogram at a certain height, is equivalent to finding a tangible gap on the height of the forestogram for a natural grouping. We propose to cut the forest when biclusters have the tendency of concentration about a center.

Assume a grid bicluster $\mathcal{C} = \mathcal{C}^{\text{row}} \times \mathcal{C}^{\text{col}}$ and therefore $\mathbf{Y}_{n\times p} \mid \mathcal{C}$ is clustered into several row and column clusters. Obviously, the total number of biclusters is $|\mathcal{C}| = |\mathcal{C}^{\text{row}}||\mathcal{C}^{\text{col}}|$. Index biclusters using $\mathbf{Y}_{IJ} = [y_{I_iJ_j}]$, where the I denotes the row cluster and J denotes the column cluster, $I =, 1 \ldots, |\mathcal{C}^{\text{row}}|$, $J = 1, \ldots, |\mathcal{C}^{\text{col}}|$, and i and j index the rows and columns of $\mathbf{Y}_{IJ}$, $i = 1, \ldots, n_I$, and $j = 1, \ldots, p_J$, respectively. Note that n_I is the number of rows in cluster

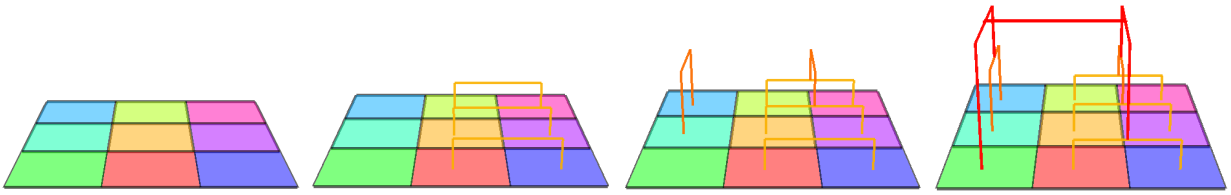


Figure 4.1 Forestogram building steps on a hypothetical 3×3 matrix. Left to right: the data matrix, merging a pair of columns, merging a pair of rows, and the completed forestogram.

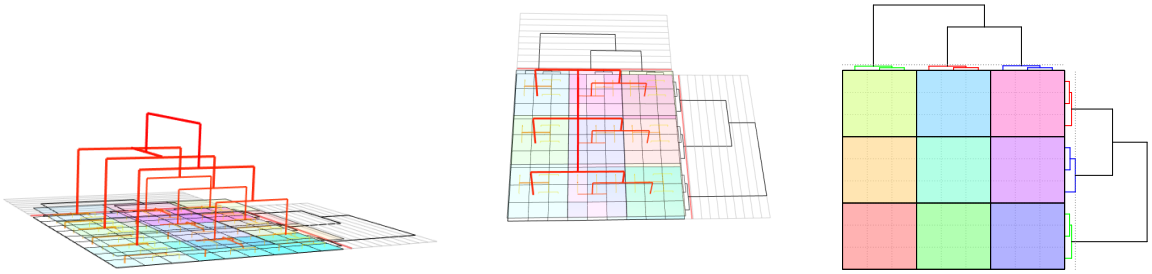


Figure 4.2 A hypothetical 9×9 matrix clustered into three row blocks and 3 column blocks after cutting the forestogram by a plane. Forestogram projection on rows and on columns provides two marginal dendrograms. Forestogram side view (left panel), above view (middle panel), projection of the forestogram on rows and columns resembling a heatmap graphics (right panel) ; the dotted horizontal and vertical lines is the projection of the cutting plane.

I, and p_J is the number of columns in cluster J,

$$n = \sum_{I=1}^{|\mathcal{C}^{\text{row}}|} n_I, \quad p = \sum_{J=1}^{|\mathcal{C}^{\text{col}}|} p_J.$$

In hierarchical clustering using a linkage, closer data have the tendency to merge together. So, it is reasonable to cut the agglomerative tree using some concentration measure. Assume the mean of data is subtracted such that the data are centered around zero. The following statistical model looks meaningful to express the concentration of bicluster IJ around a center

$$\begin{aligned} y_{IiJj} \mid \theta_{IJ} &\overset{i.i.d}{\sim} \mathcal{N}(\theta_{IJ}, \sigma^2), \\ \theta_{IJ} &\sim \mathcal{N}(0, \phi\sigma^2), \end{aligned} \tag{4.2}$$

where σ^2 can be estimated by the common within variance, and ϕ is the between-variance to within-variance ratio. We propose to cut the forestogram where this Gaussian model fits appropriately. It turns out that (4.2) yields a simple and interesting cutting strategy.

Define the within cluster variance,

$$s_{IJ}^2 = \frac{1}{n_I p_J} \sum_{i=1}^{n_I} \sum_{j=1}^{p_J} (y_{IiJj} - \bar{y}_{IJ})^2$$

and the pooled variance,

$$s^2 = \frac{1}{np} \sum_{I=1}^{|\mathcal{C}^{\text{row}}|} \sum_{J=1}^{|\mathcal{C}^{\text{col}}|} n_I p_J s_{IJ}^2.$$

The optimal number of clusters using (4.2) is found by minimizing the forest information criterion (FORIC). FORIC is a sort of penalized variance,

$$np(1 + \log 2\pi s^2) + \sum_{I=1}^{|\mathcal{C}^{\text{row}}|} \sum_{J=1}^{|\mathcal{C}^{\text{col}}|} \log(n_I p_J \phi + 1). \tag{4.3}$$

We suggest to fix $\phi = 1$ and estimate the pooled variance s^2 in each level of the forestogram. The following theorem shows how FORIC is derived.

Theorem 1 *If biclusters are generated from* (4.2),

$$-2 \log f(\mathbf{Y}) = \frac{1}{\sigma^2} \sum_{I=1}^{|\mathcal{C}^{\text{row}}|} \sum_{J=1}^{|\mathcal{C}^{\text{col}}|} n_I p_J s_{IJ}^2 + \sum_{I=1}^{|\mathcal{C}^{\text{row}}|} \sum_{J=1}^{|\mathcal{C}^{\text{col}}|} \log(n_I p_J \phi + 1). \tag{4.4}$$

Suppose the biclustering $\mathcal{C}$ is given. Following the analysis of variance notation, the Gaussian model (4.2) can be re-written in terms of a linear model, by putting the data matrix $\mathbf{Y}_{n \times p}$ in a long vector $\mathbf{y}_{np \times 1}$. The binary design matrix $\mathbf{X}_{np \times |\mathcal{C}|}$ consists of bicluster membership indicators, and $\boldsymbol{\theta}_{|\mathcal{C}| \times 1} = [\theta_{IJ}]$

$$\mathbf{y} \mid \boldsymbol{\theta} \sim \mathcal{MN}(\mathbf{X}\boldsymbol{\theta}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\theta} \sim \mathcal{MN}(\boldsymbol{\tau}, \boldsymbol{\Omega}),$$

where $\boldsymbol{\tau} = \hat{\boldsymbol{\theta}}$ that is the maximum likelihood estimator of $\boldsymbol{\theta}$, $\boldsymbol{\Omega} = \phi \mathbf{J}_1^{-1}$, and $\mathcal{MN}$ denotes the multivariate normal distribution. The conditional density is

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \right\}$$

$$\log f(\mathbf{y}|\boldsymbol{\theta}) = \log f(\mathbf{y}|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \left.\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top \left.\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$$

$\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is equal to zero at $\hat{\boldsymbol{\theta}}$ the maximum likelihood estimator of $\boldsymbol{\theta}$.

$$\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp -\frac{1}{2}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\theta})$$

$$\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} = \frac{\partial \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X}$$

and the prior distribution on $\boldsymbol{\theta}$ is

$$f(\boldsymbol{\theta}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Omega}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\tau})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}-\boldsymbol{\tau})\right\}.$$

The predictive distribution can be derived by integrating out $\boldsymbol{\theta}$ with respect to its prior

$$\begin{aligned} f(\mathbf{y}) &= \int f(\mathbf{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \exp\left\{\log f(\mathbf{y} \mid \boldsymbol{\theta})\right\} f(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int f(\mathbf{y} \mid \hat{\boldsymbol{\theta}}) \exp\left\{\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top \left(-\mathbf{X}^\top \boldsymbol{\Sigma}^{-1}\mathbf{X}\right)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} f(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \frac{f(\mathbf{y} \mid \hat{\boldsymbol{\theta}})}{\sqrt{|2\pi\boldsymbol{\Omega}|}} \int \exp\left\{\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top \left(-\mathbf{X}^\top \boldsymbol{\Sigma}^{-1}\mathbf{X}\right)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\tau})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}-\boldsymbol{\tau})\right\} d\boldsymbol{\theta} \end{aligned} \tag{4.5}$$

By taking $\boldsymbol{\tau} = \hat{\boldsymbol{\theta}}$ and the predictive distribution simplifies to

$$\begin{aligned} f(\mathbf{y}) &= \frac{f(\mathbf{y} \mid \hat{\boldsymbol{\theta}})}{\sqrt{|2\pi\boldsymbol{\Omega}|}} \int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top \left(\mathbf{X}^\top \boldsymbol{\Sigma}^{-1}\mathbf{X} + \boldsymbol{\Omega}^{-1}\right)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \\ &= \frac{f(\mathbf{y} \mid \hat{\boldsymbol{\theta}})}{\sqrt{|2\pi\boldsymbol{\Omega}|}} \sqrt{|2\pi \left(\mathbf{X}^\top \boldsymbol{\Sigma}^{-1}\mathbf{X} + \boldsymbol{\Omega}^{-1}\right)^{-1}|} \end{aligned} \tag{4.6}$$

Suppose $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is the Fisher information. In this case the Fisher information is a diagonal matrix with elements $n_I p_J$, $\mathbf{J} = \text{diag}\{n_I p_J\}$.

Model (4.2) implies $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, and $\boldsymbol{\Omega} = \phi\mathbf{J}_1^{-1}$, where $\mathbf{J}_1^{-1} = \sigma^2\mathbf{I}$ is the Fisher information

of a single observation. This setting simplifies the predictive distribution further and gives

$$f(\mathbf{y}) = \frac{f(\mathbf{y}|\hat{\boldsymbol{\theta}})}{\sqrt{|2\pi\sigma^2\phi\mathbf{I}|}}\sqrt{\left|2\pi\text{diag}\left\{\frac{\sigma^2\phi}{n_I p_J\phi+1}\right\}\right|}$$
$$= \frac{f(\mathbf{y} \mid \hat{\boldsymbol{\theta}})}{\sqrt{\prod_{I=1}^{|\mathcal{C}^{\text{row}}|}\prod_{J=1}^{|\mathcal{C}^{\text{col}}|}(n_I p_J\phi+1)}}.$$

Thus

$$-2\log f(\mathbf{y}) = -2\log f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + \sum_{I=1}^{|\mathcal{C}^{\text{row}}|}\sum_{J=1}^{|\mathcal{C}^{\text{col}}|}\log(n_I p_J\phi+1). \tag{4.7}$$

But $f(\mathbf{y} \mid \boldsymbol{\theta})$ is a Gaussian likelihood so deriving (4.4) is straightforward. Substituting σ^2 with its empirical estimator s^2 simplifies (4.7) even further and gives (4.3).

Note that (4.4) is exact, but AIC and BIC are asymptotic approximations. Using the asymptotic argument similar to BIC, we can define an extended version of FORIC,

$$-2\log f(\mathbf{Y}) \approx -2\log f(\mathbf{Y} \mid \hat{\boldsymbol{\theta}}) + \sum_{I=1}^{|\mathcal{C}^{\text{row}}|}\sum_{J=1}^{|\mathcal{C}^{\text{col}}|}\log(n_I p_J\phi+1). \tag{4.8}$$

FORIC is the adaptation of the AIC (Akaike, 1973) and BIC (Schwarz, 1978) for biclustering. Suppose biclusters are balanced and each bicluster contains $n_I = n_0$ rows, and $p_J = p_0$ columns. The extended version (4.8) coincides with the AIC if $\phi = \frac{e^2-1}{n_0 p_0}$, and coincides with the BIC if $\phi = \frac{n_0 p_0 - 1}{n_0 p_0}$.

4.3.4 Separable Biclusters

Hierarchical clustering algorithms are prone to converge to a sub-optimal grouping, due to their intrinsic greedy behavior. Here we show that a separable bicluster will appear always on the forestogram. This property holds for all linkages. Before defining a separable bicluster, we need to define the *diameter* and the *margin* concepts.

Take the submatrix $\check{\mathbf{Y}} \subset \mathbf{Y}_{n\times p}$. Denote the row and column extension of $\check{\mathbf{Y}}$ using $\check{\mathbf{Y}}^{\text{row}}$ and $\check{\mathbf{Y}}^{\text{col}}$ respectively, such that $\check{\mathbf{Y}}^{\text{row}} \cap \check{\mathbf{Y}}^{\text{col}} = \check{\mathbf{Y}}$, see Figure 4.3. Let $\mathbf{y}_i$ be the ith row of $\mathbf{Y}$ and $\mathbf{y}_j$ be the jth column of $\mathbf{Y}$. Likewise, let $\check{\mathbf{y}}_i^{\text{row}}$ be the ith row of $\check{\mathbf{Y}}^{\text{row}}$ and $\check{\mathbf{y}}_j^{\text{col}}$ is the jth column of $\check{\mathbf{Y}}^{\text{col}}$. The row margin of $\check{\mathbf{Y}}$ measures the pessimistic row-wise distance of $\check{\mathbf{Y}}^{\text{row}}$

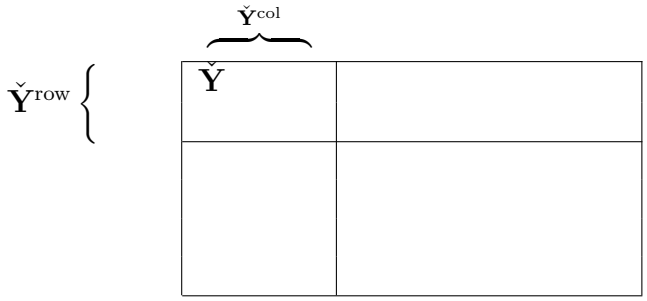


Figure 4.3 Visual illustration of submatrix $\check{\mathbf{Y}} \subset \mathbf{Y}$, extended on rows $\check{\mathbf{Y}}^{\text{row}}$, and on columns $\check{\mathbf{Y}}^{\text{col}}$.

from $\mathbf{Y}$, similarly the column margin measures the column-wise distance of $\check{\mathbf{Y}}^{\text{col}}$ from $\mathbf{Y}$,

$$\mathfrak{M}^{\text{row}} = \min_{i \neq i'} \|\check{\mathbf{y}}_i^{\text{row}} - \mathbf{y}_{i'}\|^2,$$
$$\mathfrak{M}^{\text{col}} = \min_{j \neq j'} \|\check{\mathbf{y}}_j^{\text{col}} - \mathbf{y}_{j'}\|^2.$$

Definition 1 *: Margin of bicluster* $\check{\mathbf{Y}}$ *is the minimum of row and column margins,*

$$\mathfrak{M}(\check{\mathbf{Y}}) = \min\left\{\mathfrak{M}^{\text{row}}, \mathfrak{M}^{\text{col}}\right\}.$$

Define the diameter of $\check{\mathbf{Y}}$ using row diameter and column diameter

$$\mathfrak{D}^{\text{row}} = \max_{i \neq i'} \|\check{\mathbf{y}}_i^{\text{row}} - \check{\mathbf{y}}_{i'}^{\text{row}}\|^2,$$
$$\mathfrak{D}^{\text{col}} = \max_{j \neq j'} \|\check{\mathbf{y}}_j^{\text{col}} - \check{\mathbf{y}}_{j'}^{\text{col}}\|^2,$$

Definition 2 *: Diameter of bicluster* $\check{\mathbf{Y}}$ *is the maximum of row and column diameters*

$$\mathfrak{D} = \max\left\{\mathfrak{D}^{\text{row}}, \mathfrak{D}^{\text{col}}\right\}.$$

Separability of a bicluster is defined by putting a condition on its margin and diameter.

Definition 3 *: Bicluster* $\check{\mathbf{Y}}$ *is separable if* $\mathfrak{M} > \mathfrak{D}$.

In the following theorem we study the relationship between separability and forestogram.

Theorem 2 *Separable submatrix* $\check{\mathbf{Y}}$ *always appear on the forestogram, regardless of the chosen linkage.*

The proof is by contradiction. Here we only concentrate on rows i.e. supposing $\mathfrak{D} = \mathfrak{D}^{\text{row}}$ and $\mathfrak{M} = \mathfrak{M}^{\text{row}}$, and only focus on the complete bilinkage, but the argument is equally valid for other cases.

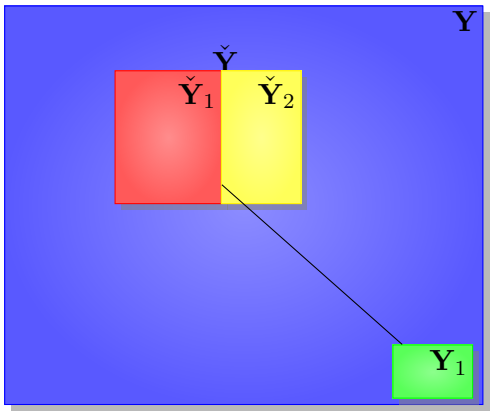


Figure 4.4 Notation for a *separable bicluster* $\check{\mathbf{Y}} \subset \mathbf{Y}$.

Suppose bicluster $\check{\mathbf{Y}} \subset \mathbf{Y}$ is separable, Figure 4.4 helps to follow the notation. From the separability we know $\mathfrak{M} > \mathfrak{D}$. Now assume $\mathbf{Y}_1 \subset \mathbf{Y}$ is merged with $\check{\mathbf{Y}}_1 \subset \check{\mathbf{Y}}$, at a certain step, before $(\check{\mathbf{Y}}_1, \check{\mathbf{Y}}_2)$ merge together, $\mathbf{Y}_1 \cap \check{\mathbf{Y}} = \emptyset$, and $\check{\mathbf{Y}}_1 \cup \check{\mathbf{Y}}_2 = \check{\mathbf{Y}}$. Suppose $\mathbf{y}_{1i}$ denotes the rows of $\mathbf{Y}_1$, $\check{\mathbf{y}}_{1i}$ denotes the rows of $\check{\mathbf{Y}}^{\text{row}}$, and $\check{\check{\mathbf{y}}}_{2i}$ denotes the rows of $\check{\check{\mathbf{Y}}}_2^{\text{row}}$, for some $\check{\check{\mathbf{Y}}}_2 \subset \check{\mathbf{Y}}_2$. By the definition of complete linkage, merging $\mathbf{Y}_1$ with $\check{\mathbf{Y}}_1$ means

$$\max_{i \neq i'} \|\mathbf{y}_{1i} - \check{\mathbf{y}}_{1i'}\| < \max_{i \neq i'} \|\check{\check{\mathbf{y}}}_{2i} - \check{\mathbf{y}}_{1i'}\|, \tag{4.9}$$

and by definition of diameter

$$\max_{i \neq i'} \|\check{\check{\mathbf{y}}}_{2i} - \check{\mathbf{y}}_{1i'}\| < \mathfrak{D}. \tag{4.10}$$

From (4.9) and (4.10)

$$\max_{i \neq i'} \|\mathbf{y}_{1i} - \check{\mathbf{y}}_{1i'}\| < \mathfrak{D},$$

which turns out to be a contradiction, because by separability of $\check{\mathbf{Y}}$

$$\min_{i \neq i'} \|\mathbf{y}_{1i} - \check{\mathbf{y}}_{1i'}\| > \mathfrak{D}.$$

Theorem 2 states that separable biclusters are kept intact during the hierarchical agglomeration. Such biclusters are recovered by cutting the forestogram at a specific level.

4.4 Computational Complexity

A brute-force implementation of forestogram is of time complexity order $\mathcal{O}(n^3+p^3)$. This price is expensive for moderate matrices, and restricts the algorithm applicability on *omics* data. We provide computational tricks to improve the time complexity of the algorithm.

Hierarchical clustering algorithms use a dissimilarity matrix to store the result of computation in an $n \times n$ matrix where n is the number of rows. The algorithm takes the advantage of avoiding process of the pairwise dissimilarities repeatedly, by augmenting the stored data. One may prefer to compute the dissimilarities *on fly* to avoid storing the dissimilarity matrix. However, on-fly computation saves the storage, with the price of increasing the computation. In the following we adapt the Lance-Williams property (Lance and Williams, 1966) into hierarchical biclustering implementation to accelerate the computations.

4.4.1 Lance-Williams Speed-up

For each merge at each level of hierarchical clustering, a dissimilarity matrix for each pair of clusters is required. After each merge, the dissimilarity for newly merged clusters must be updated. Lance and Williams (1966) developed a concise formula to use the previous distance information, to update the dissimilarity matrix.

Suppose the merging cluster is denoted by $\mathcal{C}_1 \cup \mathcal{C}_2$, and $\mathcal{C}$ denotes another disjoint cluster in the same level of hierarchy

$$D(\mathcal{C}_1 \cup \mathcal{C}_2, \mathcal{C}) = \delta_1 D(\mathcal{C}_1, \mathcal{C}) + \delta_2 D(\mathcal{C}_2, \mathcal{C}) + \delta_3 D(\mathcal{C}_1, \mathcal{C}_2) + \delta_4 |D(\mathcal{C}_1, \mathcal{C}) - D(\mathcal{C}_2, \mathcal{C})|.$$

Table 4.2 gives more details about the coefficients $\delta_i, i = 1, \ldots, 4$.

4.4.2 Time Complexity

The implementation of hierarchical biclustering requires identifying the closest two clusters. The search algorithm looks up n times on the row dissimilarity matrix and p times on the column dissimilarity matrix. However, the dissimilarity matrix is shrunken after each merge, thus the overall computational complexity is $\sum_{i=1}^{n} i^2 + \sum_{j=1}^{p} j^2$ which is of order $\mathcal{O}(n^3 + p^3)$. A proper implementation of the Lance-Williams technique speeds up the algorithm to $\mathcal{O}(n^2 + p^2)$.

Table 4.2 Lance-Williams coefficient merge updates for different linkages, if the Euclidean distance defines the linkage.

Linkage	δ_1	δ_2	δ_3	δ_4
Single	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Average	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Centroid	$\frac{\|\|\mathcal{C}_1\|\|}{\|\|\mathcal{C}_1\|\|+\|\|\mathcal{C}_2\|\|}$	$\frac{\|\|\mathcal{C}_2\|\|}{\|\|\mathcal{C}_1\|\|+\|\|\mathcal{C}_2\|\|}$	$-\frac{\|\|\mathcal{C}_1\|\|\,\|\|\mathcal{C}_2\|\|}{(\|\|\mathcal{C}_1\|\|+\|\|\mathcal{C}_2\|\|)^2}$	0
Ward	$\frac{\|\|\mathcal{C}_1\|\|+\|\|\mathcal{C}_1\cup\mathcal{C}_2\|\|}{\|\|\mathcal{C}_1\|\|+\|\|\mathcal{C}_2\|\|+\|\|\mathcal{C}_1\cup\mathcal{C}_2\|\|}$	$\frac{\|\|\mathcal{C}_2\|\|+\|\|\mathcal{C}_1\cup\mathcal{C}_2\|\|}{\|\|\mathcal{C}_1\|\|+\|\|\mathcal{C}_2\|\|+\|\|\mathcal{C}_1\cup\mathcal{C}_2\|\|}$	$-\frac{\|\|\mathcal{C}_1\cup\mathcal{C}_2\|\|}{\|\|\mathcal{C}_1\|\|+\|\|\mathcal{C}_2\|\|+\|\|\mathcal{C}_1\cup\mathcal{C}_2\|\|}$	0
Median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0

4.4.3 Space Complexity

The amount of memory required to run the algorithm is another important factor. If an $n \times p$ data matrix fits in the computer memory, the algorithm must reserve extra space for computation and storing the dissimilarity matrices. Hierarchical biclustering uses two dissimilarity matrices, and stores all pairwise dissimilarities for rows and columns. Therefore, in early steps of the algorithm, all pairwise distances are computed and initiated in two different matrices, an $n \times n$ matrix for row dissimilarity and an $p \times p$ matrix for column dissimilarity. Using the Lance-William property, only a row group and a column group will be altered at each iteration of the algorithm. This implies $\mathcal{O}(n^2 + p^2)$ for the space.

In the following, we investigate our efficient implementation on a synthetic matrix of data by fixing the number of columns to 10, and varying the number of rows. In a similar setting rows are fixed to 10 and the number of columns is varied. Figure 4.5 confirms the quadratic complexity in term of rows and columns. This speed up implementation is available in the beta version of the `R package` released in Ghaemi *et al.* (2017c).

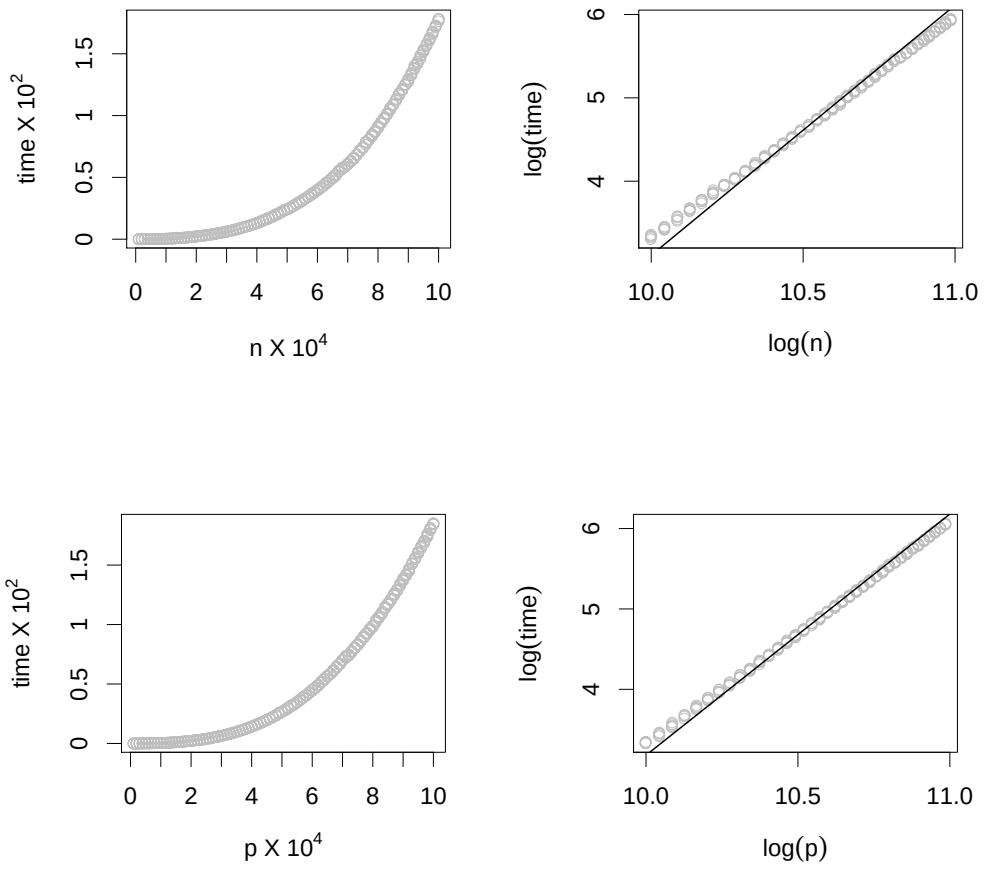


Figure 4.5 Time required to build the forestogram as the number of rows n increase (top panels), and as the number of columns p increase (bottom panels). The top right panel confirms that the algorithm is quadratic in n, the bottom right panel confirms that the algorithm is quadratic in p; the solid line is $y = \beta_0 + 2x$.

4.4.4 Parallel computing for dissimilarity measure

Euclidean Distance Matrix (EDM), is an important measure in several diverse research fields such as bioinformatics, machine learning, statistics, industrial engineering, etc. and also it has a close connection to semi-definite matrices that attracts more attention because of the applicability and other theoretical aspects. Here, we revisit the mathematical definition of EDM with extending the relation to the matrix operation directly. Then, by mapping the computation to the matrix notation, we can simply scale the overload of computing to the

standard matrix toolboxes that support parallel computing especially for general purpose Graphical Processing Unit (GPU) such as Compute Unified Device Architecture (CUDA) and Open Computing Language (OpenCL).

A Euclidean distance matrix, an EDM in $\mathbb{R}_+^{n\times n}$ is an exhaustive table of distance-square d_{ij} between points taken by pair from a list of n points $\{\mathbf{x}_i, i = 1, \ldots, n\}$ in $\mathbb{R}^n$; the squared metric, the measure of distance-square,

$$\mathbf{d}_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||_2^2 = (\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j) = ||\mathbf{x}_i||^2 + ||\mathbf{x}_j||^2 - 2\mathbf{x}_i^T\mathbf{x}_j$$

For $i, j = 1...n$, distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ must satisfy the definition of any *metric space* for *Euclidean metric* in $\mathbb{R}^n$ as following,

— Nonnegativity $\sqrt{\mathbf{d}_{ij}} \geq 0, \quad i \neq j$

— Self-distance $\sqrt{\mathbf{d}_{ij}} = 0, \iff \mathbf{x}_i = \mathbf{x}_j$

— Symmetry $\sqrt{\mathbf{d}_{ij}} = \sqrt{\mathbf{d}_{ji}}$

— Triangle inequality $\sqrt{\mathbf{d}_{ij}} \leq \sqrt{\mathbf{d}_{ik}} + \sqrt{\mathbf{d}_{kj}}, \quad i \neq j \neq k$

Consequently, all elements of EDM must be symmetric, nonnegative, with zero diagonal entries, where the fourth property provides upper and lower bounds for each element to conform the metric space.

Furthermore, suppose the matrix $\mathbf{Y} = \mathbf{X}^T\mathbf{X} = (\mathbf{x}_i^T\mathbf{x}_j)$, known as the *Gram matrix* of the points $\mathbf{x}_i, \ldots \mathbf{x}_N$, then, for any $\forall i, j \in \{1, \ldots n\}$, we have,

$$\mathbf{d}_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||_2^2 = ||\mathbf{x}_i||^2 + ||\mathbf{x}_j||^2 - 2\mathbf{x}_i^T\mathbf{x}_j = \mathbf{y}_{ii} + \mathbf{y}_{jj} - 2\mathbf{y}_{ij}$$

This can be further extended in the full matrix operation to rewrite the EDM by,

$$\mathbf{D} = \mathtt{diag}(\mathbf{Y})\mathbf{e}^T + \mathbf{e}\mathtt{diag}(\mathbf{Y})^T - 2\mathbf{Y}$$

where $\mathbf{e} \in \mathbb{R}^n$ is the vector of all ones.

This parallel representation of computing the dissimilarity measure is the bottleneck of computing when Lance-Williams speed-up is not feasible. For example minimax linkage cannot be written in the form of Lance-Williams formula (Bien and Tibshirani, 2011). Or in the case of big data, if the dissimilarity matrix does not fit the memory the pairwise similarity has to recomputed at each iteration. Thus, in the light of parallel computation of dissimilarity measure, one can use the EDM algorithm for accelerating the computing.

4.4.5 R-package

The forestogram implementation for computational visualized framework is available in the context of `R` package shown in Figure 4.6. This package consists of two components, the engine is implemented in `C`, and the interface is developed in `R` based on the `RGL` library. The engine computes the hbiclust object with fairly efficient performance that is now available on `R-Forge`, http://hbiclust.r-forge.r-project.org, which can also be installed by the following `R` command,

install.packages("hbiclust", repos="http://R-Forge.R-project.org")

The interface is capable of illustrating the 3D representation of forestogram and its 2D plot counterpart after invoking the engine who is responsible for building the hbiclust object.

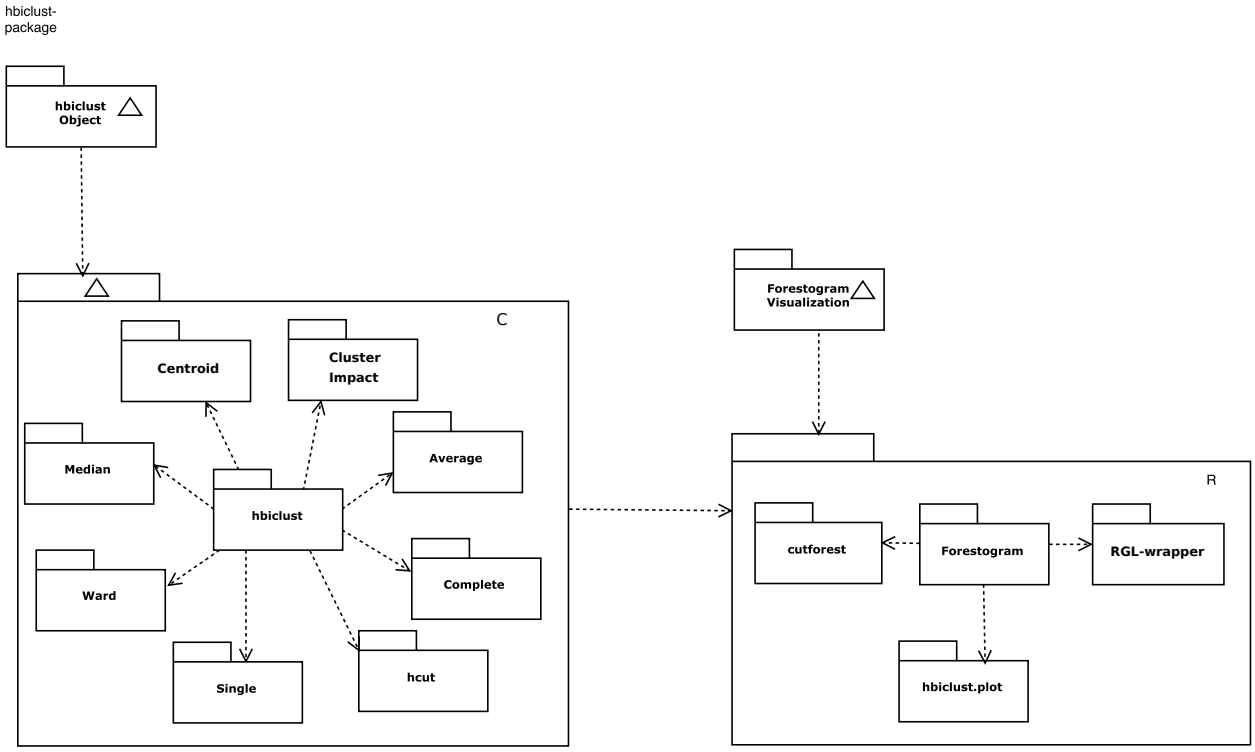


Figure 4.6 R-package architecture consists of two components, the engine is implemented in `C`, and the interface is developed in `R` based on the `RGL` library.

— `cutforest`, splits a hierarchical biclustering object, e.g., as resulting from `hbiclust`, into several row-wise groups and column-wise groups either by specifying the desired number of biclusters or the cut height.

— `hb`, hierarchical biclustering object constructed by `hb` function.

  - `k`, an integer scalar with the desired number of biclusters.
  - `h`, a numeric scalar with height where the forest should be cut, e.g., `hb$hcut`.
- `hbiclust`, this function biclusters a data matrix into a set of blocks of row-column clusters in a bottom-up hierarchical manner. The Bayesian model viewpoint that is mapped on this structure provides an automatic cutting point to reveal the number of blocks with the corresponding label assignment using the `cutforest` function. The hierarchical block-clusters can be visualized in 2D or 3D space by calling `hbiclust.plot` and `forestogram` functions.
  - `x`, a numeric matrix, with clustering individuals on row and variables on column.
  - `method`, a string variable consisting of popular hierarchical clustering linkages, such as, "average", "centroid", "complete", "median", "single", and "ward". If nothing is declared the function sets "ward" as the default linkage, that is equivalent to the Bayesian model based biclustering where the automatic cutting point is statistically meaningful.
- `forestogram`, hierarchical biclustering produces a 3D multi-tree object called, `forestogram` that is very similar to dendrogram with an augmented orthogonal merge representing the direction of merge. This extra information is useful to illustrate the order of rows and columns merges that a specific bicluster is evolved from. `forestogram` function demonstrates this 3D object that is formed hierarchically from the merge of biclusters constructed from `hbiclust` function.
  - `hb`, a hierarchical biclustering object constructed by `hbiclust` function.
  - `cut_height`, the height where the forestogram should be cut to reveal the biclusters. The `hbiclust` function computes this height approximately with FORIC trick available in `hb$hcut` to provide a guess for the number of biclusters.
  - `draw_cut`, a binary variable to show the cut surface on the forestogram corresponding to the biclusters at the given height.
  - `draw_side_tree`, a binary variable to show the 2D dendrograms on row and column side of the forestogram.
- `hbiclust.plot`, this function projects the 3D image of forestogram onto the 2D space where corresponding row and column merges are attached as dendrograms on row and column of the matrix, respectively. This function draws a graph similar to `heatmap` except the correlation colors that are replaced by bicluster's color at cutting height.
  - `hb`, a hierarchical biclustering object constructed by `hbiclust` function.
  - `cut_height`, the height where the forestogram should be cut to reveal the biclusters. The `hbiclust` function computes this height approximately with FORIC trick available in `hb$hcut` to provide a guess for the number of biclusters.

4.5 Simulation

In order to compare hierarchical method with other common biclustering techniques we generate a square 30×30 matrix, and a rectangular 150×30 matrix, divided into three clusters on rows and three clusters on columns. Both simulations consist of 10 columns in their column clusters.

In the square setup, each row cluster includes 10 rows see Figure 4.7, but in the rectangular simulation each row cluster simulation includes 50 rows. Each of the three biclusters is generated from uniform distribution of range 1 and varying mean $(-\Delta, 0, \Delta)$. The parameter $\Delta \in \{0.5, 1.0\}$ reflects the separability of biclusters. The larger the Δ is, the more separable biclusters are.

$\overbrace{\quad}^{10}$	$\overbrace{\quad}^{10}$	$\overbrace{\quad}^{10}$	
$-\Delta$	0	Δ	$\}10$
Δ	$-\Delta$	0	$\}10$
0	Δ	$-\Delta$	$\}10$

Figure 4.7 Symmetric simulation data consist of a matrix of size 30×30 with 9 biclusters. Each bicluster contains 100 data from uniform distribution with 10 rows in row cluster and 10 columns in column clusters. The parameter Δ controls the separability of biclusters.

The joint clustering of Lazzeroni and Owen (2002) and Cheng and Church (2000) are often used as standard biclustering methods. We found Cheng and Church (2000) performed poorly, so we report the *plaid* model of Lazzeroni and Owen (2002) only. The codes for Lazzeroni and Owen (2002) and Cheng and Church (2000) are available in the R package `biclust` R package (Kaiser *et al.*, 2013). Our results are based on the implementation of Lazzeroni and Owen (2002) developed by Turner *et al.* (2005).

There are few methods that combine biclustering with a visual guide. Practitioners often use the *heatmap* to produce a visualization of joint clusters. The `heatmap` produces a visualization using independent row and column dendrograms. Convex biclustering (Chen *et al.*, 2015) implmented in the R package `cvxclustr` is a new technique with a visualization similar to dendrogram, produced by shrinking the mean of biclusters towards a common mean. Our method is released as a beta version in Ghaemi *et al.* (2017c) on `R-forge`.

We created three version of forestogram by varying the linkages. All linkages are defined using the Euclidean distance. The linkages include single, average, and Ward. Other linkages behavior was similar to the average linkage, and therefore not reported. A fully automatic version of forestogram is produced by cutting the forestogram by minimizing FORIC. The

number of biclusters for all methods is set to 9. Note that, even after fixing the number of clusters the grouping may be different. Default parameters in the R package is used for the competing methods.

Table 4.3 summarizes the performance of all techniques using the adjusted Rand index of Hubert and Arabie (1985) implemented in the R package `mclust` (Fraley and Raftery, 1999). The adjusted Rand index is bounded from below by 0, and from above by 1. It gets the upper bound if the estimated biclustering matches the true clustering. We generated 100 replications of randomly generated data sets, and run different biclustering techniques. The average of the adjusted Rand index is reported. The maximum standard error is 0.1, so all reported digits are significant.

Table 4.3 admits by increasing the separation parameter Δ the performance of all methods can be improved. Changing the matrix from square to rectangle increases the number of rows from 10 to 50. This change in data size, improves the clustering performance over column clusters, for all methods except for convex, and for heatmap single linkage.

It turns out the single linkage in heatmap implementation is an inefficient method, but the performance improves significantly after being implemented as a bilinkage. The automatic cut using FORIC on forestogram is the best for forests built using the Ward bilinkage. Plaid model appears to be the least favorable technique.

Table 4.3 The performance of different biclustering techniques using the average adjusted Rand index $\times 100$. The larger the adjusted Rand index is, the better the performance will be.

	Dimension	30×30				150×30			
	Separation	$\Delta = 0.5$		$\Delta = 1$		$\Delta = 0.5$		$\Delta = 1$	
	Side	row	col	row	col	row	col	row	col
Forestogram	Auto Single	55	55	55	55	56	100	56	100
	Auto Average	55	55	55	55	56	100	56	100
	Auto Ward	55	55	55	55	100	100	100	100
	Single	80	55	100	100	94	100	100	100
	Average	100	99	100	100	100	100	100	100
	Ward	100	99	100	100	100	100	100	100
Heatmap	Single	53	53	100	100	0	100	100	100
	Average	100	99	100	100	99	100	100	100
	Ward	100	99	100	100	100	100	100	100
Plaid	Bicluster	0	0	43	99	0	60	77	94
Convex	Bicluster	54	0	100	100	0	100	100	100

4.6 Application

The yeast galactose gene expression data (Ideker *et al.*, 2001) investigates the influence of the **gal** gene family that allows cells to consume galactose, as a source of carbon. A perturbation is made in two different ways, related to a specific pathway component: i) eliminating one of the **gal** genes or ii) a wild-type for each subject regardless of galactose existence. We consider a sub-matrix of this data, widely analyzed by other researchers. The analysis of the entire data set 3935×20 is feasible thanks to the computational acceleration of the algorithm. For a similar analyses see Yeung *et al.* (2003); Yeung and Ruzzo (2001); Fowler and Heard (2012).

Each value in this data matrix is an average of four replicates. We cluster $\log_{10}$ of data with no preprocessing. The data are available in the supplementary material of Ideker *et al.* (2001). Forestogram helps to recognize similar group of genes with the same reaction to the genetic perturbation. Figure 4.8 (bottom panel) is the two-dimensional projection of the forestogram corresponding to Figure 4.8 (top panel). Presence of gene **gal** perturbation is indicated by + sign.

The **gal4** is the only gene that stays in the same cluster regardless of whether galactose present or not after perturbation. This means the presence or absence of galactose has no effect on **gal4**. A similar result is reported in Fowler and Heard (2012) but with a Bayesian biclustering model.

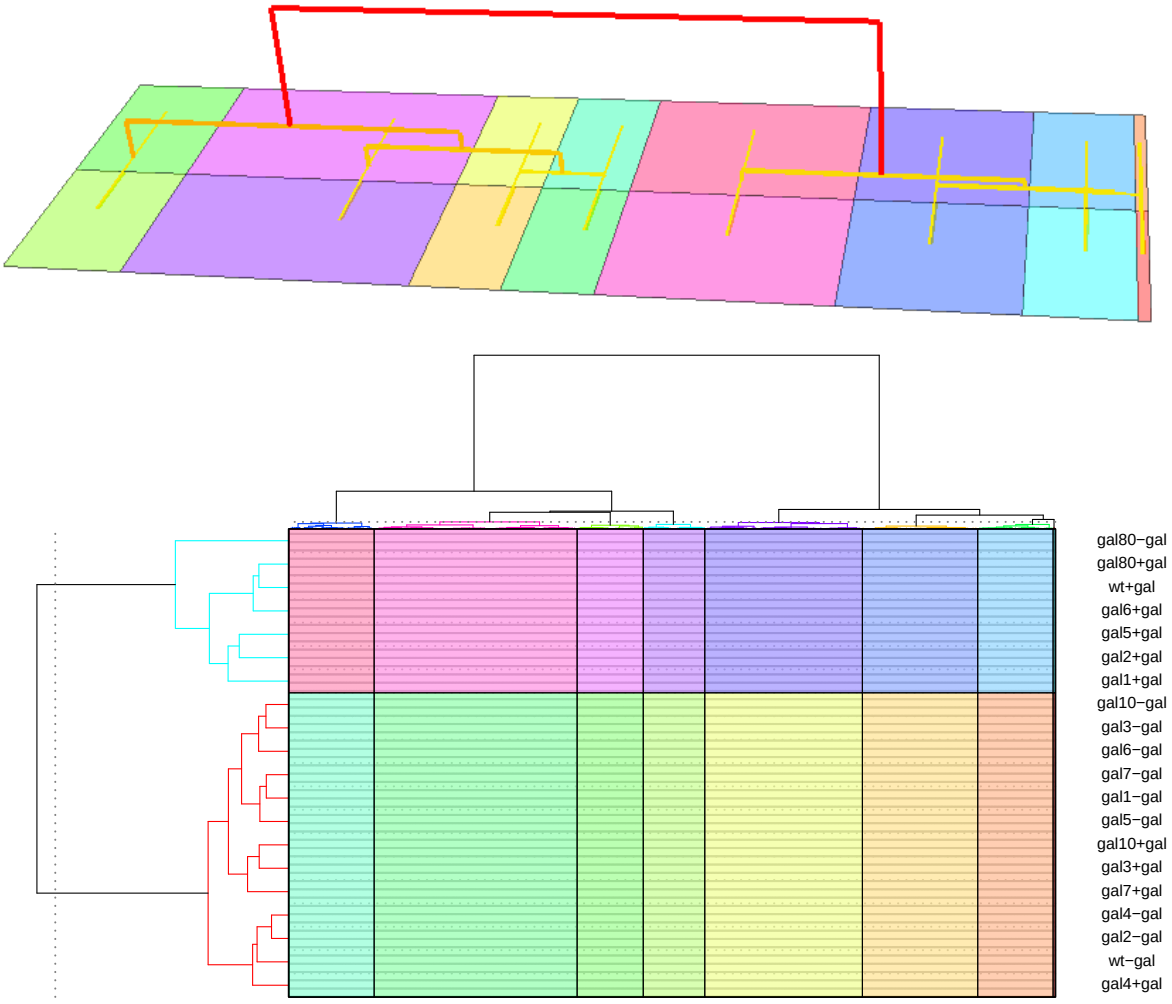


Figure 4.8 Top panel: forestogram produced using Ward bilinkage with automatic cut using FORIC. Bottom panel: two-dimensional projection of forestogram on rows and columns.

CHAPTER 5 ARTICLE 2: A VISUAL SEGMENTATION METHOD FOR TEMPORAL SMART CARD DATA

5.1 Abstract

In many cities, worldwide public transit companies use smart card system to manage fare collection. Analysis of this collected information provides a comprehensive insight of user's influence in the interactive public transit network. In this regard, analysis of temporal data, describing the time of entering to the public transit network is considered as the most substantial component of the data gathered from the smart cards. Classical distance-based techniques are not always suitable to analyze this time series data. A novel projection with intuitive visual map from higher dimension into a three-dimensional clock-like space is suggested to reveal the underlying temporal pattern of public transit users. This projection retains the temporal distance between any arbitrary pair of time-stamped data with meaningful visualization. Consequently, this information is fed into a hierarchical clustering algorithm as a method of data segmentation to discover the pattern of users. [1]

Keyword clustering ; public transit ; smart card ; temporal pattern ; projection.

5.2 Inrtoduction

Public transit serves the society to solve their mobility in almost every country (Gallotti and Barthelemy, 2015). Thus, inter-disciplinary challenges of public transit is attended in several branch of science and engineering (Gkiotsalitis and Stathopoulos, 2015). Progress of smart data and the use of automated payment system provides a rich source of data, whose its analysis can promote the economy (Weisbrod and Reno, 2009), reduce the air pollution, and enhance the quality of life (Ma *et al.*, 2015). In this regard, diverse combination of tools and techniques from various disciplines, e.g. data mining, machine learning, urban computing, urban planning, management, business, civil engineering, industrial engineering, statistics, mathematical engineering, geographic information system (GIS), and high-performance computing are vital to extract the meaningful piece of information from such data.

In most of public transit studies, bus stops and subway stations play the central role, regardless of the temporal features of the data. The frequency of the used locations is utilized to construct a model for identifying the user behavior. This knowledge is helpful to provide particular services in each station or bus stop. Nonetheless, such models are incapable to

1. MS. Ghaemi, B. Agard, M. Trépanier, V. Partovi Nia. "A visual segmentation method for temporal smart card data", *Transportmetrica A: Transport Science*, vol. 13, no. 5, pp. 381–404, 2017.

uncover the detailed behavioral pattern of users. In most of recent researches summary statistics such as the frequency of travel days, the count of similar starting boarding times, the number of similar transit sequences, and the repetition of similar stop/station sequences are extracted as descriptive features to be fed into clustering algorithms with few justification and explanatory translation. In recent years, user satisfaction from public transit system, quality of service and perceived quality of bus transit model are investigated based on reliability, length of journey, and driver amiability (Bordagaray *et al.*, 2014; Del Castillo and Benitez, 2013; de Oña and de Oña, 2015).

Despite the extensive research that has been done on public transit domain, various obstacles arise for specific purposes. Such specific purposes require particularly new computationally efficient views to address them. In this study, the problem of user clustering is attacked. The ultimate aim is to uncover the temporal behavior of users in their monthly trips.

The aim of this research work is to identify group of similar users relying on the gathered data from smart cards. More specifically, groups of similar user focusing on temporal aspect of the smart card data are identified. To this end, in Section 5.4 we propose a projection technique which is able to transform a vector of hourly usage associated to each smart card into a three dimensional feature vector that lays out the hidden temporal patterns. Accordingly, we deploy a hierarchical clustering algorithm to elicit the coherent internal representation of users in terms of analogous temporal behavior. In Section 5.6 experimental results of one month record of smart card data from Gatineau (a city in western Québec, Canada) is analyzed to illustrate the effectiveness of our suggested technique.

5.3 State-of-the-art

5.3.1 Recent research papers on the analysis of smart card data

Public transit systems have been expanded independently in many cities regardless of their size. Thus, having a strategic plan of Integrated Smart Card Fare Collection System (ISFCS) is necessary in development and enlargement of the public transit network. ISFCS fills the gap of different public transit operators and better meets the passengers' needs and satisfaction. Barriers of ISFCS and their possible solutions are discussed in Yahya and Noor (2008). In Pelletier *et al.* (2011) several other aspects of ISFCS are considered from technologies to privacy issues in three levels of management including, strategic, tactical, and operational. Moreover, discussion and comparison of planning, scheduling, and survival modeling for many different purposes are provided in Pelletier *et al.* (2011).

Describing user behavior in public transit network is one of the main issues that can

be revealed via the smart cards data (Ma *et al.*, 2013). Accordingly, finding a measure to evaluate and disclose behavioral patterns from the history of user's habits is a crucial part of Smart Card Fare Collection System (SCFCS) analysis. Various measures are proposed in Morency *et al.* (2006) by considering the variability of users' behavior over smart card data of ten months. Agard *et al.* (2008) applied k-means on weekly boarding ; this study permitted to identify large temporal users' behavior and detect important changes in users' schedule (spring break for example) but the optimal number of clusters is difficult to identify and new estimation techniques for determining the number of clusters in such data seem necessary. In Lathia and Capra (2011), two viewpoints are investigated to measure the transportation system's performance ; self-report of users' feedback, and real behavior while they are encouraged by various incentives. Lathia and Capra (2011) and Herrera *et al.* (2010) concluded that smart card data is as important as human activity on mobile phone data for designing future infrastructure and guidance of travelers. Therefore, human mobility could be modeled according to the smart card data as one of the big data sources concerning human activity.

Smart card data contain worthwhile digital information of daily locations visited at certain period by a large number of individuals. Besides, other source of information can be combined with this data such as mobile phone, GPS tracker vehicle, e.g. bike, car, motorcycle, credit card transactions, social network, (Herrera *et al.*, 2010; Gkiotsalitis and Stathopoulos, 2015). This helpful information could be utilized to characterize and model urban mobility patterns (Hasan *et al.*, 2012; Järv *et al.*, 2014). Other useful information such as travel time and number of passengers for the sake of congestion analysis and planning improvement could be possibly extracted as well (Fuse *et al.*, 2012).

Kusakabe and Asakura (2014) proposed a data fusion approach in order to estimate the trip purpose and then interpret the observed behavioral features. They are able to successfully distinguish the following different reasons: (a) commuting, (b) leisure or business and (c) returning home in 86,2% of their available trip data.

Ma *et al.* (2013) used a data mining technique to understand regular travel patterns in Beijing, China. First they constructed trip chains, then extracted regular patterns using clustering that leads to specific trip rules.

Ali *et al.* (2016) analyze electronic fare transactions for analyzing travel behavior of the users, in Seoul, South Korea. They used an open-source agent-based transport simulation package, `MATSim`, over smart card data to model input demand. This study permitted to generate micro-simulation travel demand models.

Amoung others, Trépanier *et al.* (2007) and Alsger *et al.* (2016) explore smart card data in order to estimate trip's origin and destination. Origin of trips are relatively easy to define, thanks to the first boarding check, but destinations may require prediction.

Data representation in public transit is more complicated than conventional data sets in data mining or machine learning (Nantes *et al.*, 2016). Summary of steady sequential time model in a discrete structure is the main reason that makes it difficult to analyze the temporal behavior (Shekhar *et al.*, 2015). The focus of this research, is to deal with the temporal datasets, that could be categorized as temporal snapshot model in spatial-temporal data as in Shekhar *et al.* (2015). Most of the research works in this domain perform the data mining techniques on transformed spatial-temporal attributes in a conventional way. However, because of the intrinsic structure of spatial-temporal data, independent and identically distributed (i.i.d.) observations cannot be assumed for this sort of data. Consequently, conventional data analysis algorithms often fail to capture the essential knowledge from the data. Moreover, the extracted information has no real interpretation for the experts. These are the two principal reasons that reflect the urge of why advanced techniques are required to be tailored for public transit data.

Machine learning methods are often divided into supervised, and unsupervised sub-fields ; semi-supervised methods have attracted more attention recently. Most learning methods seek for dividing data into sub-populations. The difference between supervised and unsupervised method is the existence of training data (Hastie *et al.*, 2009). More precisely, when an indicator variable is available for sub-population allocation, the problem is called supervised learning. If dividing the whole spontaneous data into k homogeneous sub-populations is required without any guide, the problem is called unsupervised learning. Note that even the number of sub-populations, k, may be unknown.

Smart card data, may provide two distinct information: spatial and temporal. Spatial data consists of coordinates of the bus stop, e.g. latitude and longitude that could be GPS data or relative location coordinates. Temporal data describes the boarding time. According to this information, analyzing users behavior is divided into three categories, 1) Spatial patterns, 2) Temporal patterns and 3) Spatial-temporal patterns.

1. Spatial pattern analysis focuses on location, such as the bus stop information. It turns out measuring the behavioral pattern only depends on the location of bus stops taken by the users, rather than knowing the starting hour of their trip.
2. Temporal methods seek the information pertinent to the time associated to the public transit usages. Consequently, computing user's similarity score is carried out, by assuming that the bus stop information is unavailable. Thus taking the public transit at a specific time, plays the central role in this approach.
3. The third scenario, is a mixture of the spatial and temporal approaches, called spatial-temporal data analysis. It could be viewed as a combination of the last two steps or an independent new approach to deal with spatial-temporal behavioral patterns.

5.3.2 Extraction of users' temporal patterns in transportation

The extraction of users' temporal behaviors may be of value for planners. It may help them to plan the service more effectively. Classical approach to to discover users' need includes counting passengers on board. Then, a generic demand is estimated. Smart card data permit to get more information: it is possible to extract generic behavior for all the users, or it is possible to follow a specific card. Clustering algorithms permit to subdivide the whole population of users in different groups that share certain behavior. The number of groups may vary depending on the accuracy needed by planners.

The majority of clustering algorithms can be divided into distance-based methods or model-based methods. Distance-based techniques are easy to understand and simple to implement. On the contrary, model-based approaches are flexible and adapt to complex data patters, but are counter-intuitive to implement or interpret.

Hierarchical clustering is a breakthrough in distance-based clustering context, because of producing a visual guide in the form of a binary tree, known as *dendrogram*. In addition it requires little prior knowledge, except for a dissimilarity measure. The dissimilarity measure is a positive semi-definite symmetric mapping of pairs of groups onto the set of real numbers. This measure, however, may not satisfy the triangle inequality unlike a distance. Hierarchical algorithms require a dissimilarity measure to merge clusters in order to build a nested structure of clusters. The common dissimilarities include single linkage (or nearest neighbors), complete linkage (or farthest neighbors), average linkage, and centroid linkage. There are two variants of hierarchical clustering depending on the direction of the construction of the nested groups. Agglomerative clustering starts with every observation as a singleton and consequently merges the closest clusters to end up with all data in one cluster. Divisive algorithms, on the contrary, start with all data in one cluster and split the clusters until finishing with all singletons.

The nested groups generated using a hierarchical clustering algorithm of data, are visualized through a dendrogram. It provides an informative representation and visualization for different potential data structures, specifically while real hierarchical relations exist in the data. Dendrogram illustrates the nested structure or the evolutionary pattern of the members of a particular set. The idea of the dendrogram first appeared in evolutionary biology, and then applied in practice as an illustrative clustering tool in Sneath (1957). The height of the dendrogram expresses the dissimilarity between each pair of clusters. The initial groups are the leaves and every merge of clusters appears with an increasing height.

An automatic cutting point on the dendrogram has been a well-known problem for decades. Estimating a grouping, cutting a dendrogram, and model selection are closely related concepts. The ultimate estimated grouping is found by cutting the dendrogram at some

height. One expects a visible gap in the height of the dendrogram for a natural grouping, but providing a universal cutting point on a dendrogram is counter-intuitive. An approximate model selection criteria such as AIC Akaike (1973) or BIC Schwarz (1978) can be applied to cut the dendrogram if a statistical model is used to produce nested clusters (Heller and Ghahramani, 2005; Heard *et al.*, 2006). Most of the statistical models for clustering are a sort of mixture model (McLachlan *et al.*, 2004). The R package `NbClust` (Charrad *et al.*, 2014) provides 30 different techniques to discover the optimal number of clusters in a data set. However, dendrogram itself provides a fairly well description of the clusters, so that it enables the experts in each domain to have a profound insight where to cut the dendrogram for finding the appropriate groups of data.

Data mining approach is used to understand passenger's temporal behavior to exploit the interpretable clusters (Mahrsi *et al.*, 2014). This approach helps transportation operators to become aware of the customers' demands. In addition, it enables them to maintain their services and meet the user's requests more effectively. The real dataset from the metropolitan area of Rennes (France) with four weeks of smart card data containing trips of both bus and subway is tested. Furthermore, the cluster of similar temporal passengers extracted based on their boarding time, according to a generative model-based clustering approach. After, the effect of distribution of socioeconomic characteristics on the passenger temporal clusters are investigated in this study.

As another example Ortega-Tong (2013) studied the extensive database of Oyster Card transactions obtained from London's public transit users. This database is deployed to classify users based on the temporal variability, the spatial variability, the socio-demographic characteristics, the activity patterns, and the membership type. Improving the planning and the design of market research are the aims of this work, when selecting groups of homogeneous people is also of interest. Four groups of users including, regular users consist of workers and students commuting during the week, portion of them who make leisure journeys during the weekends, occasional users containing leisure travelers, and visitor travelers for tourism and business affair are investigated in this work.

Smart card data gathered from Brisbane, Australia is another source of information studied in Kieu *et al.* (2014) for strategic transit planning according to the individual travel patterns. Origins and destinations of cardholders is defined as travel regularity, and the definition of habitual time is the regular time of trips. Thus, mining the travel regularity of the frequent users could be inferred to extract the travel pattern and its purposes. Reconstruction of user trips is made by spatial and temporal characteristics, then the frequent users are grouped by applying k-means clustering technique on the trip features including, origin and destination, number of transfers, travel mode and route uses, total travel time, and transfer

time. In the last step, three level of Density Based Spatial Clustering of Application with Noise (DBSCAN) are applied to find the travel regularity Kieu *et al.* (2014).

Schedules are a proper solution for the public transit user and for the public transit service provider. Most of the time, service providers operate on the same schedule in weekdays from Monday to Friday, and maintain distinct schedules for Saturdays and Sundays, assuming that the public transit user follows the same travel behavior during weekdays. It could be true for people with a regular schedule. However, society is constantly changing and more people now work only four days while other people work distantly once or twice a week. In addition, there are an increasing number of citizens with non-regular schedule such as immigrants or tourists. Hence, it becomes more of interest of the service provider to measure and predict the amount of regularity of public transit users, through their time-stamped smart card transaction database. By applying learning methods on smart card database we aim to divide the users into several sub-populations to obtain the clusters of users according to their behavior. These clusters can be put back in the context of daily mobility. Hopefully, by the analysis of these clusters we better understand the categories of the users, especially those who have a regular pattern of travel (Morency *et al.*, 2010).

5.3.3 Synthesis and justification of the needs

The state of the art shows large interest of researchers in extracting knowledge from smart card data. Authors propose many directions, tools, and methods to explore this rich source of data. Those contributions may be classified in three main domains

The first set of studies focuses on understanding the data, e.g. what happens on the network ? this aspect is about extracting many indicators, evaluate characteristics and identify behaviors in the data. All information available from the smart cards are manipulated, formatted, and analyzed boarding times, stops, lines and directions are the main information that are explored here.

The second set of studies deals with explaining the behaviors, e.g. why do we observe those behaviors ? Here researchers explore the reasons that explain what they observe. Various sources of external data are widely used, depending on the intention of the authors. The idea is to cross, fusion, and predict from a data set. The smart card data are put in relation with the external sources of data to explain why one behavior or another is observed.

The third set of research consists of taking advantage of the extracted knowledge to help in decision making. Various objectives are considered i) to improve the service for the user, with no supplementary cost for the transit operator, ii) to keep the same service but with minimized cost for the transit system operator.

Nevertheless, all of those topics rely on a good extraction of the user's behavior from

the smart card data set. Besides, that extraction could be improved considering a better metric for the comparison of users' behavior. Traditional metrics consider Euclidean distance (which could not be used in our case), but also Dynamic Time Warping (DTW) and cross correlation. The two last metrics are powerful and widely recognized for the comparison of time series, but specific properties required for the analysis of customers' behavior could be used to improve this extraction. The main goal of this paper is to contribute in this aspect.

5.4 Proposed methodology

We suggest a two-stage visual method for analysis of temporal user behavior. The first stage consists of semi-circle projection to reduce the high-dimensional data into highly interpretable lower space. In the second stage a hierarchical clustering is applied on the preprocessed data to extract the cluster structure for the expert.

We offer a simple mapping of boarding time information to the Cartesian coordinates. This suggestion is a sort of a multidimensional scaling (Borg and Groenen, 2005), when some equalities and inequalities are proposed for certain distance between individuals. The mapping, that we call *Semi-Circle Projection* (SCP) is easier to understand in the polar coordinate, i.e. in terms of radius and angle.

First, reserve the center of a half circle for zero boarding time. For users with one boarding, take radius equal to $r_1 = 1$ and move the angle from 0 to π depending on the time of boarding. For vectors with 2 boardings, take radius $r_2 = 2$. Generalization for users with sequence of n boardings is then straightforward. Choose $r_n = n$ and move the angle according to the average time of boardings. However, the identity function $r_n = n$ diverges for large n. Choice of a converging r_n helps us to renormalize the half circles for long binary sequences, if needed. Our suggestion is $r_n = (1 + \frac{1}{n})^n$ having $\lim_{n\to\infty} r_n = e$, where e is the Euler constant. The third coordinate is required, as this method maps [0 1 1 0] and [1 0 0 1] on the same location after projection, because both have the same number of unit values (being two) and both have the same average of positions for unit values (being 2.5). This appeals to add another coordinate with a scale measure over the position of the unit values to distinguish these two users from each other in the projected space. We suggest the standard deviation of the position of the unit values as the z coordinate, giving a larger value to [1 0 0 1] comparing to [0 1 1 0], so they would not be mapped on the same point in three-dimensional projection. Suppose there are m user-day entities, organized in the binary matrix $X_{m\times L}$ whose rows indicate the daily usage for specific smart card. This mapping can be formalized as follows,

$$\theta_{m\times L} = \begin{bmatrix} 1 & 2 & \cdots & L \\ 1 & 2 & \cdots & L \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \cdots & L \end{bmatrix} \odot X$$

where $\odot$ is Hadamard (elementwise) product operator. Let r represents the number of boardings, thus for all smart cards with equal r, reduced data in the new space is written as

$$\left[x_i = r_i \sin\left(\frac{\pi}{Ln_i}\sum_{j=1}^{L}\theta_{ij}\right),\ \ y_i = r_i \cos\left(\frac{\pi}{Ln_i}\sum_{j=1}^{L}\theta_{ij}\right), \right.$$
$$\left. z_i = \sqrt{\frac{1}{L-1}\left\{\sum_{\{j|\theta_{ij}>0\}}\theta_{ij}^2 - \frac{\left(\sum_{j=1}^{L}\theta_{ij}\right)^2}{L}\right\}} \right].$$

The number of boardings for the i'th user-day as $n_i = \sum_{j=1}^{L} X_{ij}$ that is the number of unit elements in the vector X_i, $L = 24$ denotes the number of time intervals, and $r_i = \left(1 + \frac{1}{n_i}\right)^{n_i}$.

The suggested simple transformation maps a binary sequence of any length to the Cartesian coordinates of only three dimensions. Implementing this method for traveled days, is compressed into only three dimensions, hugely facilitating further computation, analysis, and data visualization. The x coordinate represents the number of trips, the y coordinate represents the average time of trips, and the z axis shows the time variability of the trips. The result of this method on the dataset from Table 5.1 is shown in Figure 5.1.

5.5 Projection properties

The suggested projection includes two parameters, r_i, and θ_i for each temporal usage X_i that contribute to map the binary representation into three dimensional semi-circle space. The number of ones or frequency of usage is one of the most important factor which leads the projection through the mentioned parameters. Thus, the properties of this projection are somehow proportional to the total amount of usage determined by sum of ones. This implies the range of r_i, θ_i and $\mathrm{var}(\theta_i)$ are decreasing by increase in frequency of usage across time. Furthermore, it is evident for any pair of temporal usage encoded in binary vectors denoted by X_1, and X_2, where $X_1 \neq X_2$, and $n_1, n_2 \in \{1, 2\}$, SCP maps them onto distinct points in three dimensional reduced space, i.e. the projection is unique. However, this interesting property may not hold for $n_i > 2$. Below we separately state some properties of the projection in terms

of these parameters r_i, θ_i and $\text{var}(\theta_i)$. Remind that the first two axes of SCP is constructed using (r_i, θ_i) and the third axis is $\sqrt{\text{var}(\theta_i)}$.

Theorem 3 *Suppose $n_i \in \mathbf{N}$, then*

$$\frac{r_{n_i+1}}{r_{n_i}} < \exp\left\{\frac{1}{n_i+1}\right\}$$

The rate of radius growth is decreasing by increase in boarding.

$$\begin{aligned} \frac{1}{n_i+1} &< \log(1+\frac{1}{n_i}) \\ \frac{n_i}{n_i+1} &< n_i \log(1+\frac{1}{n_i}) \\ \frac{n_i}{n_i+1} &< \log(r_{n_i}) \\ -\log(r_{n_i}) &< -\frac{n_i}{n_i+1} \end{aligned} \tag{5.1}$$

$$\begin{aligned} \log(1+\frac{1}{n_i+1}) &< \frac{1}{n_i+1} \\ (n_i+1)\log(1+\frac{1}{n_i+1}) &< \frac{n_i+1}{n_i+1} \\ \log(r_{n_i+1}) &< 1 \end{aligned} \tag{5.2}$$

By adding the inequalities (5.1), (5.2) we have,

$$\begin{aligned} \log(r_{n_i+1}) - \log(r_{n_i}) &< 1 - \frac{n_i}{n_i+1} \\ \log(r_{n_i+1}) - \log(r_{n_i}) &< \frac{1}{(n_i+1)} \\ \frac{r_{n_i+1}}{r_{n_i}} &< \exp\left\{\frac{1}{n_i+1}\right\} \end{aligned}$$

It is evident that r_{n_i} is an increasing sequence of n_i. Theorem 3 states that the rate of increase of the radius in SCP is very tight as the number of boarding n_i increases.

Theorem 4 *Suppose a pair of binary vectors X_i and $X_{i'}$ of length L, both with the same number of boarding n_i, then*

$$\max|\tilde{\theta}_i - \tilde{\theta}_{i'}| \leq 1 - \frac{n_i}{2L}$$

where $\tilde{\theta}_i = \frac{1}{Ln_i}\sum_{j=1}^{L}\theta_{ij}$.

The range of angle is decreasing by increase in boarding. We start with an example where $n_i = 3$. Then $X_{\underset{i}{\text{argmin}}\, \tilde{\theta}_{n_i=3}} = [1, 1, 1, \dots]$ and $X_{\underset{i}{\text{argmax}}\, \tilde{\theta}_{n_i=3}} = [\dots, 1, 1, 1]$. Therefore, the maximum range of angle $|\tilde{\theta}_i - \tilde{\theta}_{i'}|$ for a set of temporal usages with the same boarding varies between $\min \tilde{\theta}_i$, and $\max \tilde{\theta}_i$ that is $[\frac{n_i(n_i+1)/2}{n_i L}, \frac{L(L+1)-(L-n_i)(L-n_i+1)}{2n_i L}]$ according to the definition, this implies the range of angle is shrinking by $\frac{L(L+1)-(L-n_i)(L-n_i+1)}{2n_i L} - \frac{n_i(n_i+1)}{2n_i L} = (1 - \frac{n_i}{2L})$.

Theorem 4 states that the angular range decreases as n_i increases. This property along Theorem 3 gives a vague idea about concentration of data after SCP for large n_i for the x-y coordinates of the SCP.

A similar result exists for the z-coordinate as well.

Theorem 5 *Suppose a set of all possible binary vectors with n_i number of boarding, and accordingly a set of all possible binary vectors with $n_i + 1$ number of boarding. Then the projection of such points over the z-coordinate of the SCP satisfy*

$$\min z_{n_i} < \min z_{n_i+1},$$

and

$$\max z_{n_i+1} < \max z_{n_i}.$$

Trivially if $\min z_L = \max z_L$ *because the set of possible boarding with L number of boarding includes only one member.*

The range of variance is decreasing by increase in boarding. In this section we use the alternative definition of variance that can be defined as

$$\frac{1}{n_i - 1} \sum_{i=1}^{n_i} (\theta_i - \bar{\theta})^2 = \sum_{i,j=1, i \neq j}^{n} \frac{(\theta_i - \theta_j)^2}{2(n_i - 1)^2}$$

In order to prove the decrease of variance we have to show that minimum of variance is monotonically increasing by increase in boarding while the maximum of variance is monotonically decreasing.

Minimum variance of the temporal usages is monotonically increasing by increase in boarding.

Proof by induction.

First of all, we show that our claim is true for the first step.

$$\min z_{n_i=2} = \sqrt{.5} < \min z_{n_i=3} = \sqrt{1}$$

this is true according to the definition where it occurs at $X_{\underset{i}{\text{argmin}}\, z_{n_i=2}} = [1, 1, \dots]$, and $X_{\underset{i}{\text{argmin}}\, z_{n_i=3}} = [1, 1, 1, \dots]$.

Let $s_{n_i} = \sum_{i,j=1, i\neq j}^{L} (\theta_i - \theta_j)^2$ now, we suppose that $\frac{s_{n_i}}{2(n_i-1)^2} < \frac{s_{n_i+1}}{2n_i^2}$ is true, then we want to show that $\frac{s_{n_i+1}}{2n_i^2} < \frac{s_{n_i+2}}{2(n_i+1)^2}$ also holds based on the first assumption.

$$\begin{aligned} \frac{s_{n_i}}{2(n_i-1)^2} < \frac{s_{n_i+1}}{2n_i^2} &= \frac{s_{n_i} + \sum_{i=1}^{n_i}(n_i+1-i)^2}{2{n_i}^2} \\ \frac{(n_i-1)^2}{{n_i}^2} &< \frac{{n_i}^2}{(n_i+1)^2} \end{aligned} \tag{5.3}$$

Thus by multiplying two inequalities in 5.3 we have,

$$\frac{s_{n_i}}{2{n_i}^2} < \frac{s_{n_i} + \sum_{i=1}^{n_i}(n_i+1-i)^2}{2(n_i+1)^2} \tag{5.4}$$

lemma *For any $n \in \mathbf{N}$, we have*

$$\frac{\sum_{i=1}^{n}(n+1-i)^2}{2n^2} < \frac{\sum_{i=1}^{n+1}(n+2-i)^2}{2(n+1)^2} \tag{5.5}$$

Proof,

$$\begin{aligned} 2n^3 + 5n^2 + 4n + 1 &< 2n^3 + 7n^2 + 6n \\ \frac{2n^2+3n+1}{2n} &< \frac{2n^2+7n+6}{2(n+1)} \\ \frac{n(n+1)(2n+1)}{2n^2} &< \frac{(n+1)(n+2)(2n+3)}{2(n+1)^2} \\ \frac{\sum_{i=1}^{n}(n+1-i)^2}{2n^2} &< \frac{\sum_{i=1}^{n+1}(n+2-i)^2}{2(n+1)^2} \end{aligned}$$

Thus by adding two inequalities from (5.4), and (5.5) we have,

$$\frac{s_n + \sum_{i=1}^{n}(n+1-i)^2}{2n^2} < \frac{s_n + \sum_{i=1}^{n}(n+1-i)^2 + \sum_{i=1}^{n+1}(n+2-i)^2}{2(n+1)^2}$$

That is,

$$\frac{s_{n+1}}{2n^2} < \frac{s_{n+2}}{2(n+1)^2}$$

Maximum variance of the temporal usages is monotonically decreasing by increase in boarding.

Proof by induction.

First of all, we show that our claim is true for the first step.

$$\max z_{n_i=2} = \sqrt{264.5} > \max z_{n_i=3} = \sqrt{169}$$

this is true according to the definition where it occurs at $X_{\underset{i}{\text{argmax}}\, z_{n_i=2}} = [1, \ldots, 1]$, and $X_{\underset{i}{\text{argmax}}\, z_{n_i=3}} = [1, 1, \ldots, 1]$.

Now, we suppose that $\frac{s_{n_i}}{2(n_i-1)^2} > \frac{s_{n_i+1}}{2{n_i}^2}$ is true, then we show that $\frac{s_{n_i+1}}{2{n_i}^2} > \frac{s_{n_i+2}}{2(n_i+1)^2}$ based on the first assumption.

By setting $A = s_{n_i}, B = s_{n_i+1} - s_{n_i}$ such that $B = \sum_{i=1}^{\frac{n_i}{2}}(\frac{n_i}{2} + 1 - i)^2 + \sum_{i=1}^{\frac{n_i}{2}}(L - \frac{n_i}{2} - i)^2$, and $C = L - \frac{n_i}{2} - \frac{n_i}{2} - 1$, suppose,

$$\frac{A}{2(n_i-1)^2} > \frac{A+B}{2{n_i}^2} \tag{5.6}$$

Now, we have to show that the following inequality is true.

$$\frac{A+B}{2{n_i}^2} > \frac{A+2B+C}{2(n_i+1)^2}$$

By expanding equation (5.6) we have,

$$\begin{aligned} n_i^2 A &> (n_i-1)^2 A + (n_i-1)^2 B \\ (2n_i-1)A &> (n_i-1)^2 B \\ (2n_i+1)A &> (n_i-1)^2 B + 2A \end{aligned} \tag{5.7}$$

Thus we have to show that, the following inequality is true.

$$\begin{aligned} (n_i+1)^2 A + (n_i+1)^2 B &> n_i^2 A + 2n_i^2 B + n_i^2 C \\ (2n_i+1)A &> (n_i^2 - 2n_i - 1)B + n_i^2 C \end{aligned}$$

From equation (5.7) we know that $(2n_i+1)A > (n_i-1)^2 B + 2A$, now it is sufficient to show that $(n_i-1)^2 B + 2A > (n_i^2 - 2n_i - 1)B + n_i^2 C$.

By rearranging $(n_i-1)^2 B + 2A > (n_i^2 - 2n_i - 1)B + n_i^2 C$ we should show that $A + B > \frac{n_i^2}{2} C$ holds.

$$A + B = 2\sum_{j=1}^{n_i+1}\left[\sum_{i=1}^{\frac{j}{2}}(\frac{j}{2} + 1 - i)^2 + \sum_{i=1}^{\frac{j}{2}}(L - \frac{j}{2} - i)^2\right]$$

$$\sum_{i=1}^{\frac{j}{2}}(\frac{j}{2}+1-i)^2 = \frac{\frac{j}{2}(\frac{j}{2}+1)(j+1)}{6} > \frac{j^3}{24}$$

$$\sum_{i=1}^{\frac{j}{2}}(L-\frac{j}{2}-i)^2 > \frac{j}{2}C$$

$$A+B > \sum_{j=1}^{n_i}(\frac{j^3}{12}+jC) > \frac{n_i^4}{48}+\frac{n_i^2}{2}C > \frac{n_i^2}{2}C$$

Therefore, we prove that the minimum variance of temporal usages is monotonically increasing and maximum variance of temporal usages is monotonically decreasing by increase in boarding. This implies the range of variance is also monotonically decreasing by increase in boarding and for the extreme usage where one enters the network at every single hour the minimum and maximum variance is collapsed on the same point.

Adding Theorem 5 to the other properties suggests that the data have the tendency of concentration as the number of boardings increases. So if a clustering technique is implemented after projection, one may expect one or several clusters of data with large number of boardings raised naturally as the property of the concentration of data with large number of boarding. Subsequently SCP better maps the data with small number of boardings. Therefore, from public transport perspective, we recommend clustering after SCP if a clustering of data with small number of boarding is of interest.

5.6 Experimental results

Experimental design consists of two steps to analyze the data. First of all, SCP is applied on the high-dimensional binary data to project the data into the lower dimension. Next, hierarchical clustering reveals the structure of the users where similar ones grouped together. To this end, we first show the performance of SCP method on a small synthetic example by comparing our suggested method with other standard techniques. Then we use smart card data to discover similar groups of users in Gatineau transit network.

5.6.1 Demonstration of Semi-Circle Projection (SCP)

After introducing the suggested ad-hoc SCP method, we compare it with the other state-of-the-art time series distance measurements to illustrate the properties of the SCP. This demonstrates how one can improve the drawbacks for the temporal user behavior. Two commonly used distance measures, namely, cross-correlation distance, and autocorrelation-based dissimilarity distance are used from the `TSdist` package in `R` as the base measures for this comparison.

The cross-correlation based distance measure between two numeric time series is calculated by

$$D(x,y) = \sqrt{\frac{(1-\{\text{CrossCorr}(x,y,0)\}^2)}{\sum_{k=1}^{K}(1-\{\text{CrossCorr}(x,y,k)\}^2)}},$$

where $\text{CrossCorr}(x,y,k)$ is the cross-correlation between x and y at lag k, and the sum in the denominator goes from 1 to the maximum lag say K. Autocorrelation-based dissimilarity, computes the dissimilarity between a pair of numeric time series based on their estimated autocorrelation coefficients that can be calculated as $D(x,y) = \sqrt{(\rho_x - \rho_y)^\top \Omega (\rho_x - \rho_y)}$, where ρ_x, ρ_y are the estimated autocorrelation vectors of x and y respectively, Ω is a matrix of weights, and $\top$ denotes the transpose operator (Montero and Vilar, 2014).

Table 5.1 Synthetic example of temporal data associated to 13 users and the corresponding usage during 7 hours, e.g. user 1 entered the public transit in the very early hour of day where the related index is 1.

User	1	2	3	4	5	6	7	...	24
X_1	1	0	0	0	0	0	0	...	0
X_2	0	1	0	0	0	0	0	...	0
X_3	0	0	1	0	0	0	0	...	0
X_4	0	0	0	1	0	0	0	...	0
X_5	0	0	0	0	1	0	0	...	0
X_6	0	0	0	0	0	1	0	...	0
X_7	0	0	0	0	0	0	1	...	0
X_8	1	1	0	0	0	0	0	...	0
X_9	1	0	1	0	0	0	0	...	0
X_{10}	0	1	1	0	0	0	0	...	0
X_{11}	1	0	0	1	0	0	0	...	0
X_{12}	0	0	0	0	1	1	0	...	0
X_{13}	0	0	0	0	0	1	1	...	0

The results of the three different distance measures are shown in Figure 5.2, 5.3 for the users X_1 and X_8, respectively. Then $\{X_8, X_9, X_2\}$ could be considered as the first three nearest users to the user X_1 because of the similar time behavior. All three methods indicate the user X_8 as the closest user to the user X_1 in Figure 5.2, however, X_9 is selected as the second nearest user in Figure 5.2(b) while the X_2 is selected in Figure 5.2(a), 5.2(c). Despite, the reasonable justification for the first two nearest users selected by cross-correlation distance, picking the user X_{13} as the third closest user to the X_1 violates the assumption of the temporal behavior in this dataset. Autocorrelation-based dissimilarity and the SCP measures preserve the constraints of the temporal distance for the user X_1. Next, the user X_8 is taken into

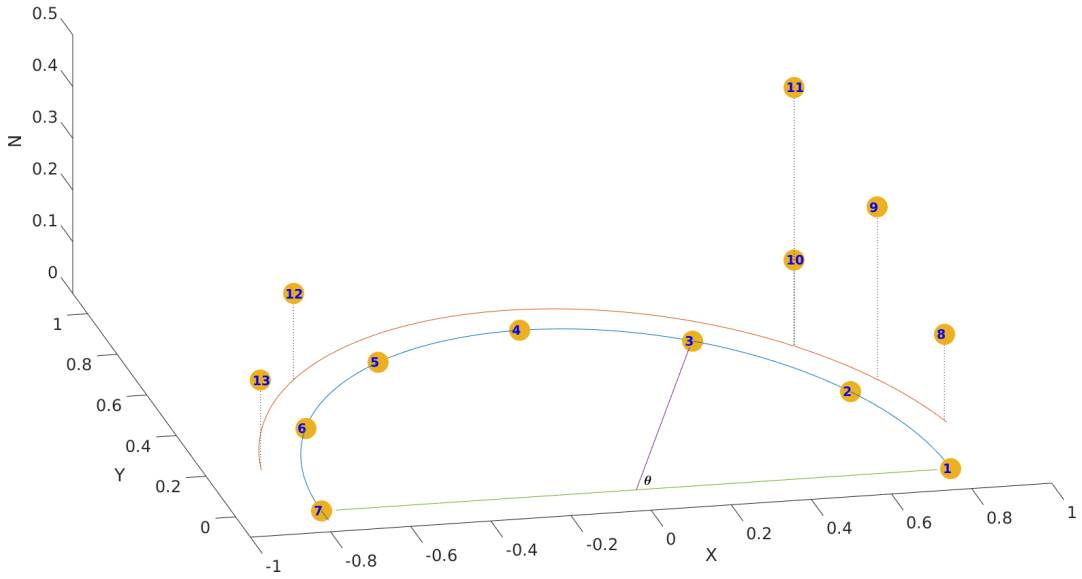
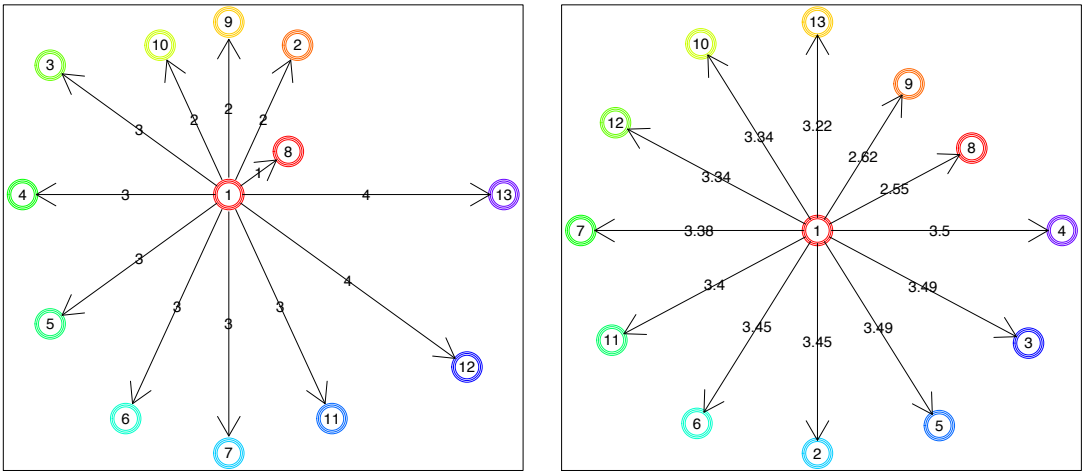


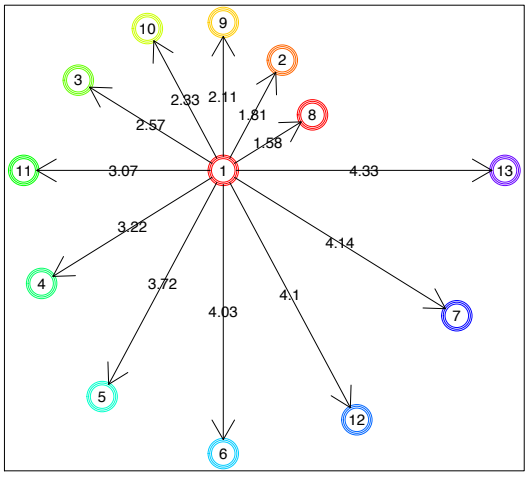
Figure 5.1 Result of the Semi-Circle Projection on the synthetic dataset from Table 5.1 in three dimension which illustrates how similar users are located close to each other.

account to follow up the performance of each method. The users $\{X_1, X_2, X_9\}$ are the first three candidates to be chosen as the nearest users to the X_8. In Figure 5.3, the selected users associated to the user X_8 are shown. Autocorrelation and the SCP are capable of picking those users as are shown in Figure 5.3(a) and 5.3(c), respectively. Yet cross-correlation is able to discover only X_1 as the second closest user while X_{13} is chosen as the first nearest similar user. Apparently, cross-correlation is not well-tailored to extract the similar users according to the temporal pattern. Regarding the discrete values of the autocorrelation distance that is redundant for couple pairs, e.g. in Figure 5.3(a), the same distance is assigned between four pairs, (X_8, X_4), (X_8, X_5), (X_8, X_6), and (X_8, X_7) which should not be the same. However, the correct order with associated distance is restrained by the SCP method. Moreover, the time series measurements are designed to give a value for a pair of vectors which requires $\binom{n}{2}$ flops. The SCP projects each data into a lower space independently to demonstrate the data in the reduced space with less computational complexity. The computational complexity of the SCP is of order $\mathcal{O}(n)$, where n is the number of projecting users. In Figure 5.1, the projected users from Table 5.1 into 3D space is shown where the aforementioned constraints are still kept.



(a) Autocorrelation

(b) Cross-correlation



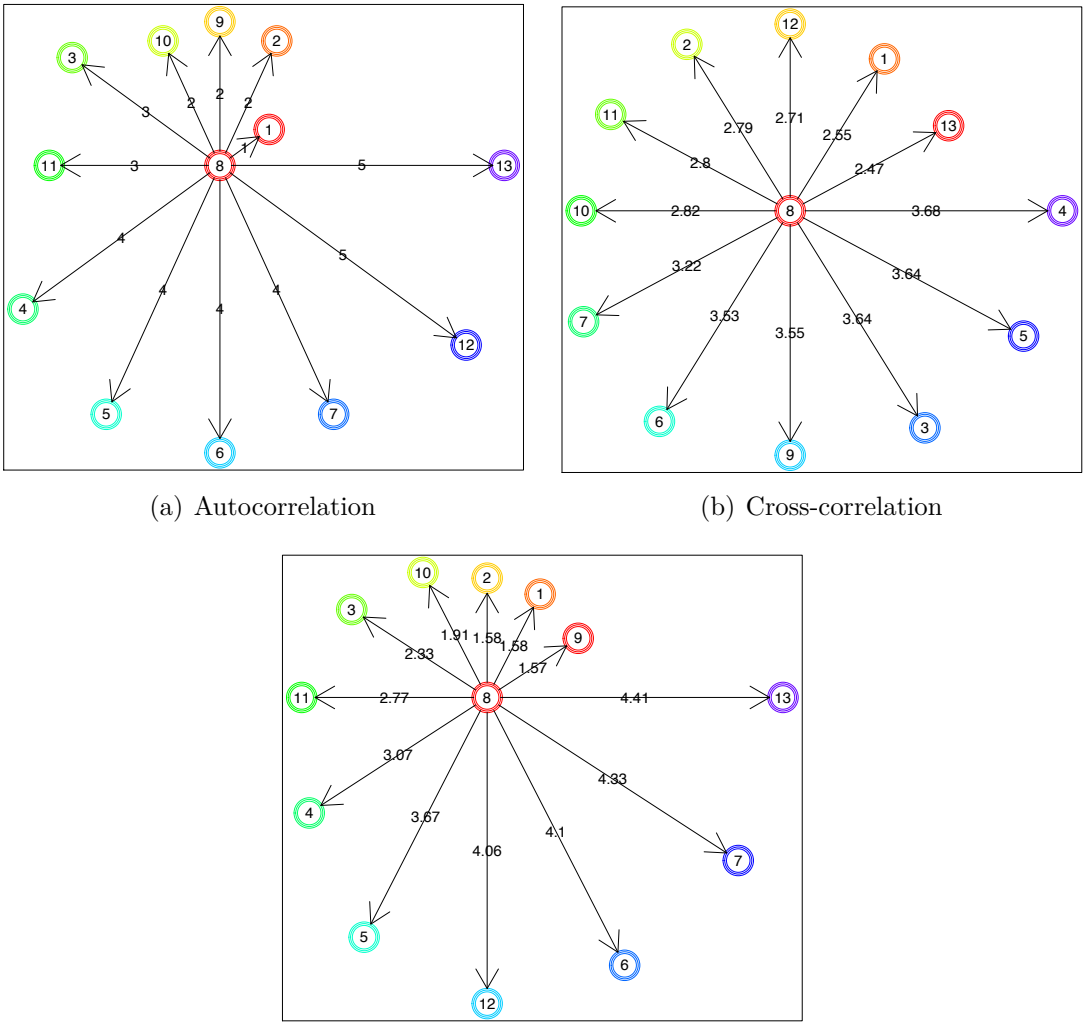
(c) Semi-circle projection

Figure 5.2 Comparison of the nearest users of X_1 with three similarity measurements, autocorrelation, cross-correlation, and semi-circle projection, respectively. As we expect, observations show that SCP method effectively sort out the similar users according to the temporal usage related to the user 1.

5.6.2 Experimenting the SCP method on Gatineau dataset

Société de transport de l'Outaouais (STO) in Gatineau, Québec, Canada, provides the data of this study. The STO authority has started to use smart card system since 2001 in its 200-buses network. Everyday, data of every transaction is gathered from public transit users at bus stops boarding passengers. For each transaction, the following properties are present:

1. Date and time of the boarding transaction ;
2. Card number and fare type ;



(a) Autocorrelation

(b) Cross-correlation

(c) SCP

Figure 5.3 Comparison of the nearest users of X_8 with three measures of similarity, autocorrelation, cross-correlation, and semi-circle projection, respectively. As it could be seen, SCP is able to find out the analogous users by projecting them into three dimensions.

3. Route number and direction;
4. Vehicle and driver numbers;
5. Stop number at boarding.

Note that for the sake of security and privacy purposes, card numbers are encrypted so that all user-information is completely anonymous. Additionally, we suggest to encode the temporal data into a $0-1$ vector whereas 24 binary vector associated to the daily hours. In this vector, occurrence of 1 at a specific index represents the usage of smart card at the corresponding hour. To deal with the binary values, discrete structure is usually suggested

as the first option to entail the data for further process.

This projection method is tested on the mid-size authority (300 buses and 220,000 inhabitants), over one month period in April 2009 (data is gathered from 753,016 transactions, with 26,176 unique users and 416,076 card-days). From the first analysis of usage histogram shown in Figure 5.4, it turns out a large subset of users prefer to take the public transit between 15-20 days per month, on average. Figure 5.5 (3D histogram of the projected users

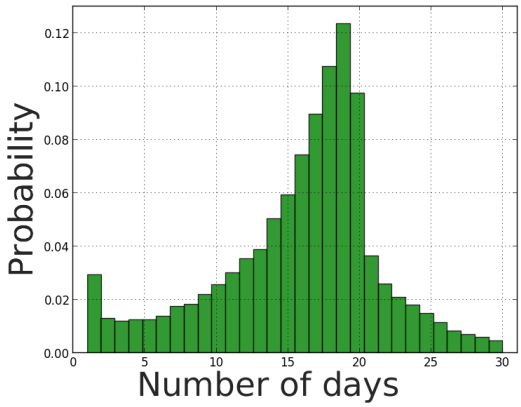


Figure 5.4 Histogram of the frequency of the traveled days in one month.

on xy-plane) demonstrates how many users overlapped on the same point without the z-axis that captures the standard deviation of the timestamps. In other words, the frequency of this histogram states the proliferation of the same sum of usage indices for different behaviors. Moreover, this illustrates the peak of the half-circle has the highest density which reflects the existence of a meaningful pattern depicted in Figure 5.5.

The dendrogram in Figure 5.6(a) shows the visual aggregation of users on the projected data. In Figure 5.6(b) existing clusters for the cutting point are illustrated. Vertical axis represents similarity measure between clusters. Similar users are grouped in the bottom of the dendrogram ; higher in the hierarchy, clusters are grouped together. The closer the groups are the more in the bottom, the more different they are bigger is the dissimilarity and higher are the steps in the dendrogram. It is then easy to identify possible cuts in the dendrogram that will stop the grouping process where too much dissimilar clusters are merged together ; it is simply about translating the red line from the top to the bottom in order to identify big steps in the dendrogram.

Different options may be possible. From Figure 5.6(a), cuts in $2, 3, 4, 6, 9$ or 18 clusters are to be considered. For a similarity/dissimilarity perspective they are close options, besides for an expert in the application domain it is possible to differentiate between these options.

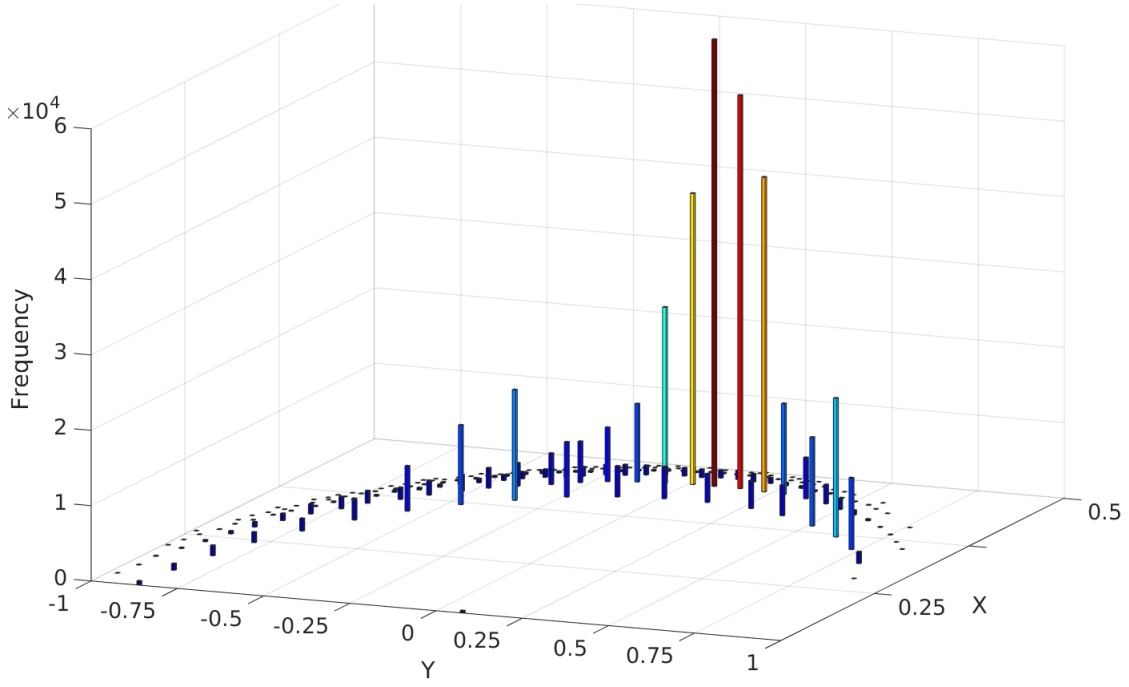


Figure 5.5 3D histogram of the overlapped projected data on xy-plane.

Considering domain expertise, a cut on the top, in $2, 3$ or 4 clusters, will be completely unbalanced (few customers on the right will be separated from all other users on the left in another group), this option will not be useful for the context of explaining users' behavior, we would have one conclusion that applies to a large group, which is not really useful for the practitioner. Options in 9 or 18 groups are still available, both could be processed and compared. From a methodological aspect here, we selected 18 groups for a more accurate prospect. These clusters of user behavior in public transit are described as the following categories.

Single trip: as it is shown in Figure 5.7, significant number of patterns belong to this group ranging from early morning to late night, though with different number of users and distribution that is shown in Figure 5.6. Members of this group commute once or few times a day for one-way monadic trips at certain hours.

Regular commuters: four types of users take the public transit regularly as can be seen in Figure 5.8. The pattern of these loyal users shows the frequent of usage all day long. Taking the number of users in this group into account, considerable portion of users are categorized as regular users who rely on public transit for their daily trip.

Late commuters: another category of users is determined in Figure 5.9, which demonstrates typical evening and late night usages for the most expeditions occurring or done on

many occasions. These users usually enter the public transit network after the work for different purposes or come back to home late night.

Long day: this category is characterized by a two-peak distribution of the transactions during a typical day of travel shown in Figure 5.10. This is generic schedule of morning and evening peak period travel time.

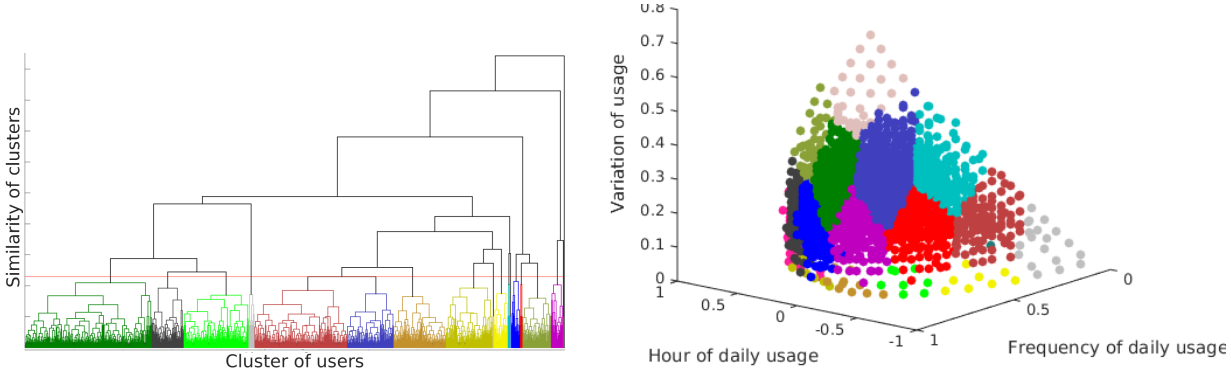
Midday versus long day: the patterns of long day users and midday travelers are shown in Figure 5.10(a) and Figure 5.10(b), respectively. The former group of users usually behave as a combination of regular users and late commuters. This reflects the fact that long day users intrinsically take the public transit for habitual commuting to work and late night circulation. In analogy, the subscriber of the latter cluster, are more similar to the late commuters whose pattern is shifter over to left. This implies the spread of transactions revolves around lunch peak time and evening rush hour.

Active versus inactive: Figure 5.11 shows the active users and the inactive smart cards' behaviors. Active users never miss any bus along their way as it could be seen in Figure 5.11(a). However, few users never used their smart card for the given one month interval with null pattern as it is shown in Figure 5.11(b). These two groups of users have the most extreme behavior in the public transit network in comparison to the remaining ones.

Let us look at the proportion of cards-day corresponding to each cluster in Figure 5.12, by working days, Saturdays and Sundays for the duration of the given month. It shows that during the working days (from Monday to Friday), the proportion of regular and single clusters (pendulum AM-PM trips) are much higher than the other ones. However, the proportion of the late commuters and the active clusters increased over the weekend, while a sharp drop of regular users is seen. This happens because people move later in the afternoon, and the trips are less characterized by pendulum movements like in the working days. It is also interesting to look at the distribution of the cluster by the entire days of month shown in Figure 5.13 where the same patterns could be found scaled by the frequency of trips per day.

5.7 Conclusion and Discussion

User's behavior modeling is crucial for predicting future financial gain, transportation scheduling, and traffic load. Thus, the main objective of the data mining on the public transit data is uncovering people's behavior. We presented the analysis of the public transit smart card transactions by projecting the high-dimensional binary vector of the temporal data into a three dimensional semi-circle and three-dimensional space. The new representation of the data provides a visual guide to a better understanding of the temporal pattern. Seventeen clusters are identified in terms of single trip, regular users, late commuters, long day, midday,



(a) Dendrogram of hierarchical clusters.

(b) Corresponding temporal patterns.

Figure 5.6 Dendrogram of the hierarchical clustering with the associated clusters of the projected data. Figure 5.6(a), shows 18 clusters, the total temporal patterns that exist for the one month period of the smart card usage. These clusters are shown on the projected data, in Figure 5.6(b).

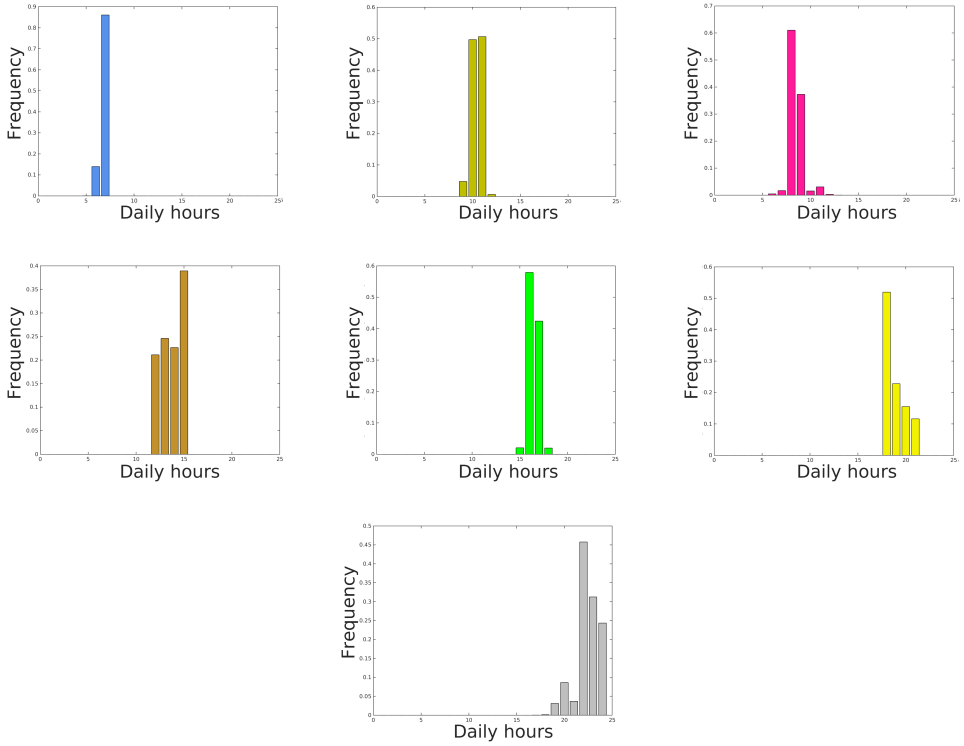


Figure 5.7 Pattern of single trips ordered by early to late.

active and inactive groups as the temporal behavior of the users by applying agglomerative hierarchical clustering on the transformed data. Despite a continuous variable carries more

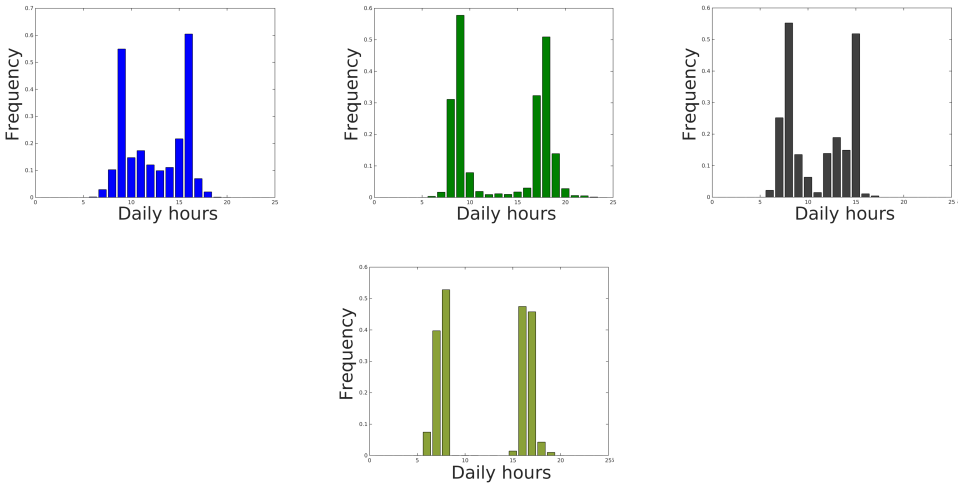


Figure 5.8 Pattern of regular users.

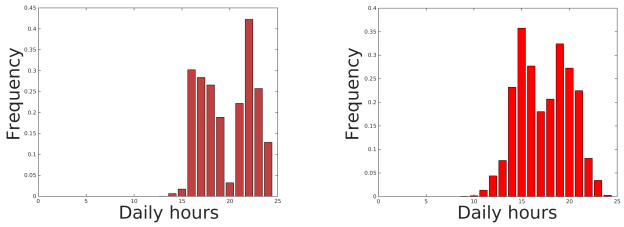
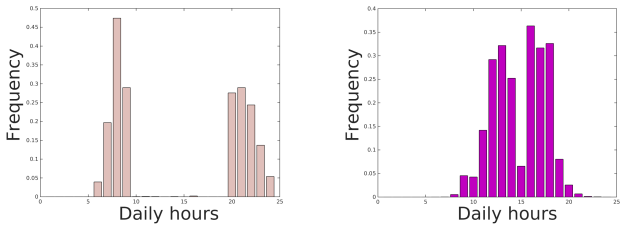


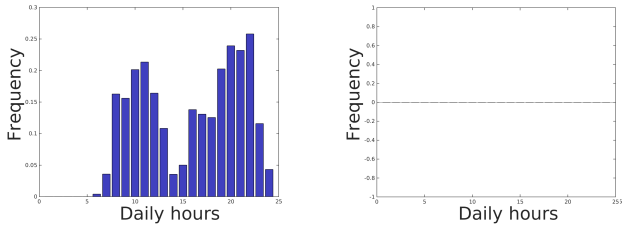
Figure 5.9 Patterns of late commuters.



(a) Long day pattern (b) Midday pattern

Figure 5.10 Patterns of long-day trips vs midday excursion.

information, binary data carries little amount of information compared to the continuous variable. This motivated us to transform a binary sequence to one or several continuous variable to execute a computationally efficient analysis. In this research study, 24 hours user-day pattern is used as the original data, however, our method is flexible to analyze even more



(a) Active users (b) Inactive smart cards

Figure 5.11 Patterns of active users versus inactive cards.

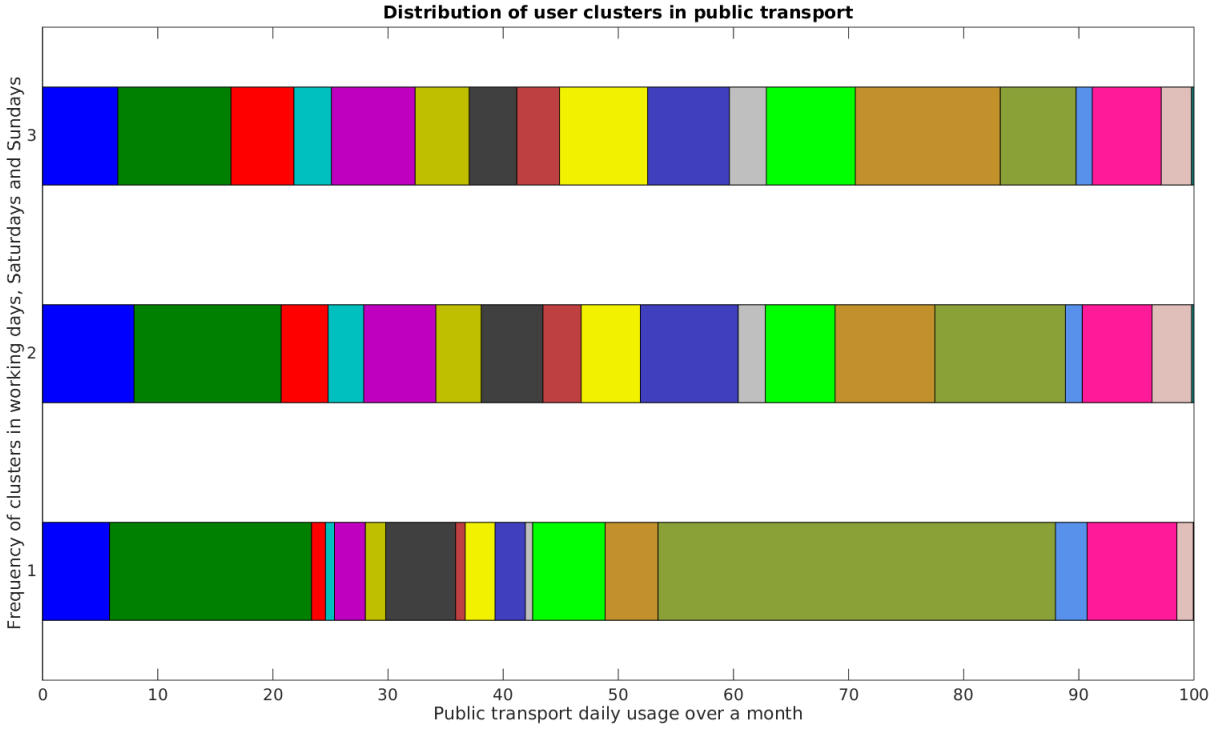


Figure 5.12 Distribution of clusters shown in Figure 5.6 for usual working days and weekends.

complicated patterns such as 30 day user-day, or 365 user-day efficiently.

Most of the data mining algorithms are developed for continuous variables that we can take the advantage of them, if we properly transform binary data to continuous and informative space. Benefiting from a proper transformation we also gain computational feasibility through dimension reduction. Developing a particular data structure, one can decrease the computational time complexity of the hierarchical clustering algorithm from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 \log n)$ or even $\mathcal{O}(n^2)$ by certain properties of the algorithm, where n is the number of users. Remembering the binary vector of length 24×30 for each individual using the public

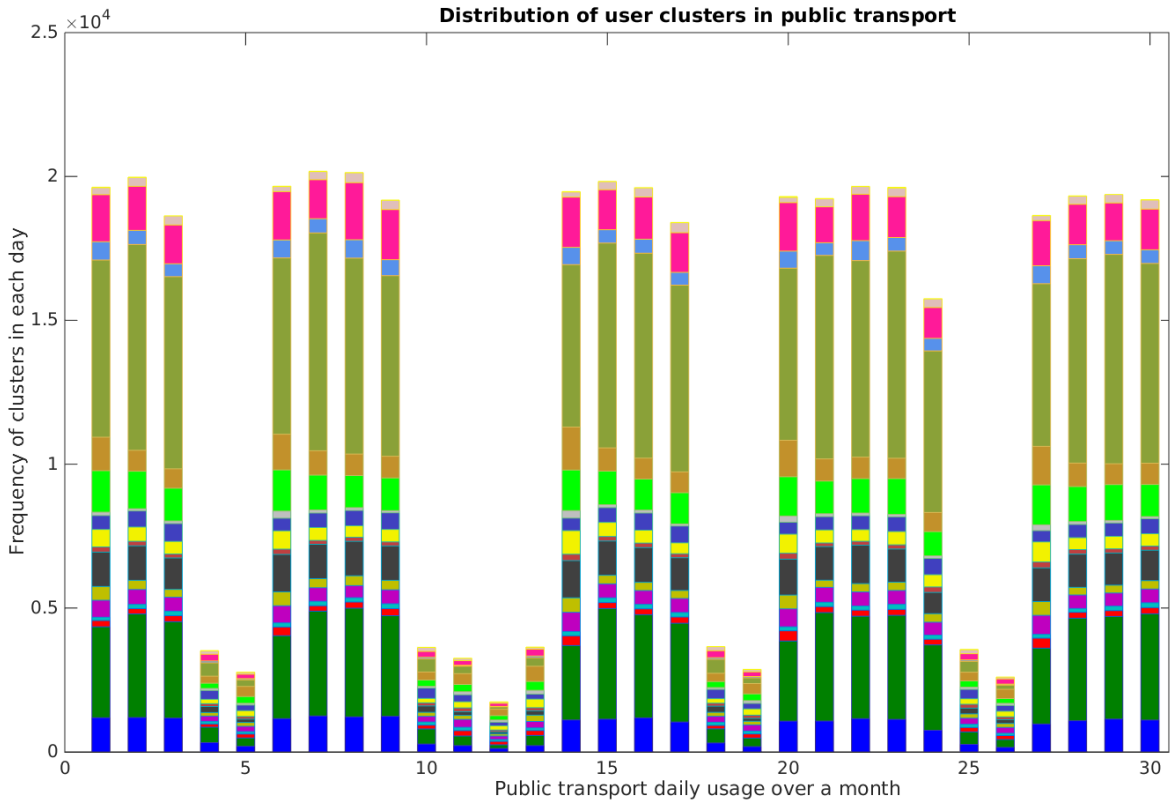


Figure 5.13 Daily cluster distribution for the entire period of the month.

transit in one month, if only 1000 people use the public transit, the amount of storage and computing facility required for analysis of such data with recent data mining algorithms is cumbersome, even with today computational power. The issue becomes worse if we analyze data of several years.

Several issues arise as future directions of this work. First, there is a need to define an equivalent metric on the binary space corresponding to the Euclidean measure on the projected three dimensional space. Second, the analysis of spatial data remains as the open question for our future research because of the existence of complex scenarios which require sophisticated techniques to compute the similarity of the users. Third, the technique can be applied to other sorts of vectors, not only including timestamps, but also the location of boarding on the territory, the route sequences, route types, etc. if the data are encoded in a binary vector.

The following subsections dealing with spatial data analysis are not part of the published journal paper. The challenges of spatial data analysis was already published as a conference paper in Ghaemi *et al.* (2015). Recently under the light of forestogram development, we

suggest a new perspective to extract the spatial patterns through temporal latent variable. In the following we present the spatial data analysis with the observed results produced by forestogram biclustering.

5.8 Challenges in Spatial Data Analysis Targeting Public Transit

Spatial data contains worthwhile information about the geographical details of each bus stop and are stored sequentially following the order of temporal usage. Although enough information about the coordinates of bus stops are available, defining a measure of similarity of behaviors in the public transport network is troublesome. The main issues about similar trips in the spatial case can be summarized into the following two questions. First, are two users similar according to the similar bus stops they usually use every day ? Second, are they categorized in the homogeneous group of users, if their resultant traversed distance resembles ? Moreover, it is possible to consider the following scenarios to realize how this spatial criterion is difficult to define.

Fig. 5.15, shows three users, red, blue, and green, who use the public transport from the same starting point and leaving the system at the same point as well, however, they use different number of trips in various directions. Hence, their resultant traversed distance is quite identical while each uses a different path. This example shows that how the answer to the two asked question can quit change the measure of user similarity in the spatial data analysis task.

Fig. 5.16, points out two red and blue users start and end their trips using the same bus stops but in the opposite directions. In contrary to the Fig. 5.15, regardless of the resultant trips, one can define the similarity only according to the bus stops. This may reflect the trip patterns of the same user who travels between home to work and vice versa in different time period.

In Fig. 5.17, it turns out that it is possible to ask even the third question. Despite, the starting points and the ending points are distinct for both users and none of them use the same bus stops, still one directional routing pattern is emerged. Besides, that it can be a consequence of taking different buses from variant inceptions to the terminations, taking the same bus stops in the same route but in different time intervals would be the other reason. The former instance, is happening in the spatial-temporal data analysis.

With the similar argument described for Fig. 5.17, Fig. 5.18, the directional routing pattern can happen in a symmetric manner as well. This symmetrical property, is held in the horizontal orientation in Fig. 5.18, vertical orientation, $x = y$, $x = -y$, and etc. are also likely to consider.

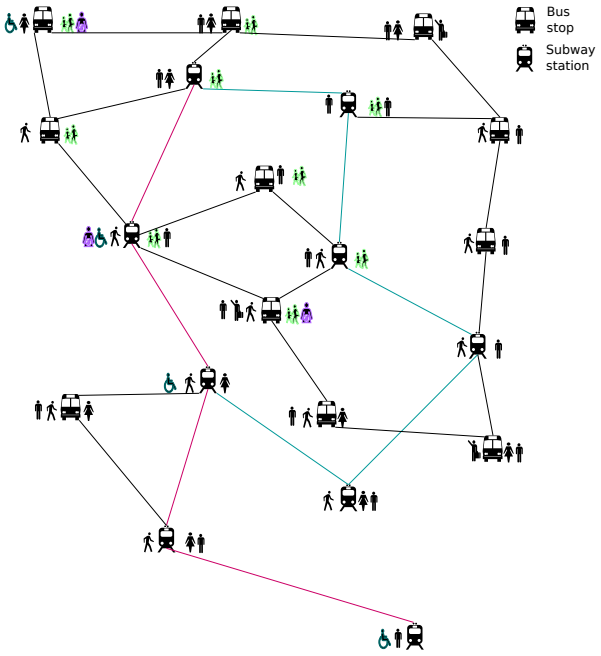


Figure 5.14 A typical network of public transport

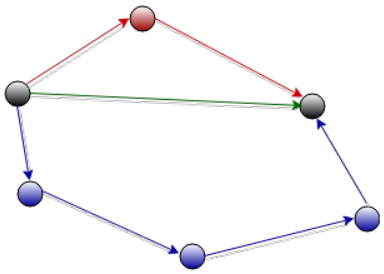


Figure 5.15 Three users with the same start point and end point

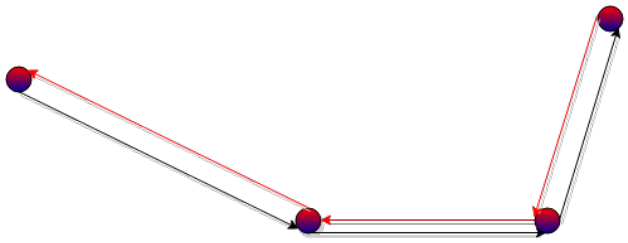


Figure 5.16 Two users taking the same buses in opposite directions

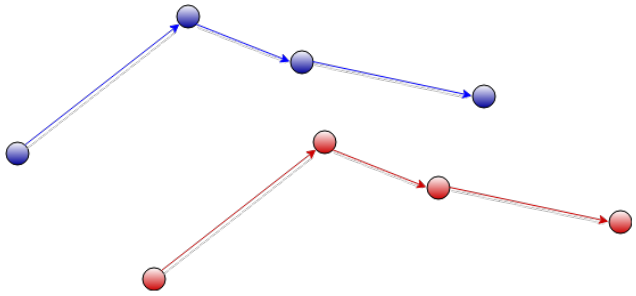


Figure 5.17 Two users with the same directional pattern

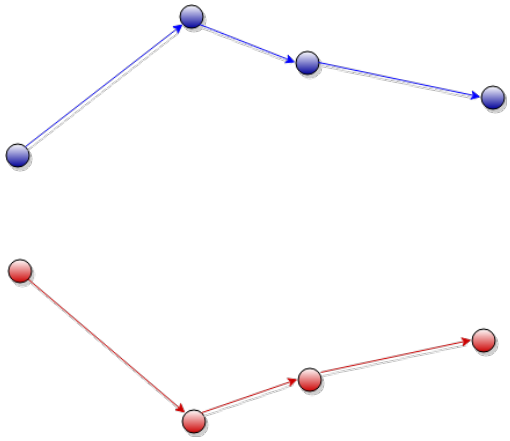


Figure 5.18 Two users with the same symmetric directional pattern

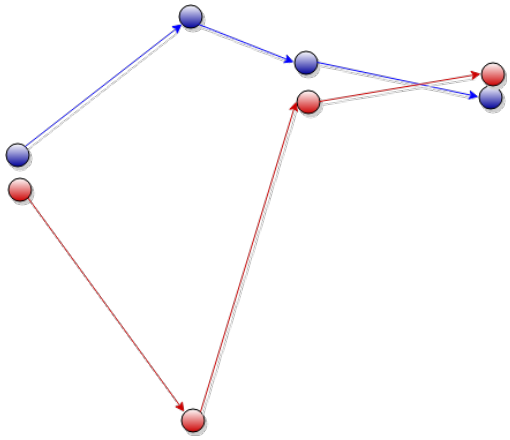


Figure 5.19 Two users with the same pattern of usage except one

Consider a case where two users are following almost the same sequence of bus stops order except one. Fig. 5.19 shows this situation, this behavior can also belong to the schedule of one user in two different days. This anomaly would be likely to occur too when frequent bus stops are used by similar users. Defining this type of usage pattern as an outlier or might be a noise, because of fault in storing or capturing devices, quite depends on the definition of user similarity criterion.

In Fig. 5.20, two users are shown, the total trip and bus stops taken by user blue, is a subset of the used bus stops by user red. In this circumstance, two users are utilizing the public transport roughly alike in a particular part of their schedule, nevertheless they behave differently beyond that interval. Hence, it turns out, the number of the taken bus stations is another important factor in defining the user similarity in the spatial domain.

Fig. 5.21, shows the other scenario, where the two users differ in the number of trips. Similar to Fig. 5.20, the blue trajectory that used bus stops, is a subset of taken bus stops by the red user. However, the resultant traversed distance is almost the same for both users. The sequence of bus stop usage, associate to the closely similar pair of users differs in the number of taken bus stops.

Suppose two users who take the same bus stops not necessarily in the same order, during their daily trip. In other words, permutations of the same bus stops can amount to the totally different resultant traversed distance. As it is shown in Fig. 5.22, the same bus stops are still shared between the two users without the same usage pattern. This often gets more complicated when temporal information gets involved in this sort of data analysis dilemma.

In other scenario, two users might use the public transport exactly in the same order, except the starting point and end point. This is an ordinary pattern that would be used by users who live in different parts of the city, they take the same bus stops during their daily trip. For instance, Fig. 5.23 shows two users following the same pattern in the downtown area, while living far away from each other.

In the former circumstances, Euclidean distance between bus stops, was assumed in the definition of the user similarity. This presumption can be violated, if the utilization of the

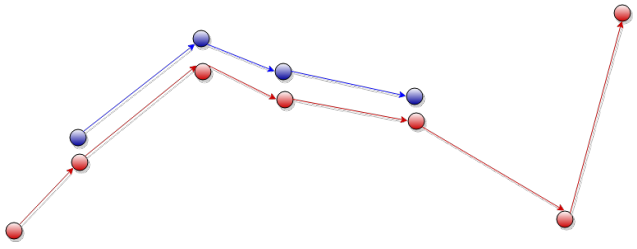


Figure 5.20 Two users with partial similarity pattern

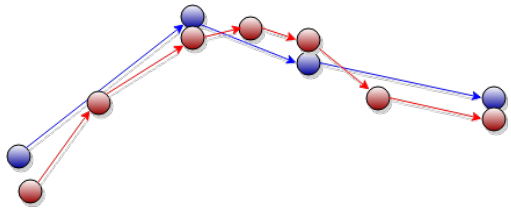


Figure 5.21 The same resultant traversed distance with different bus stops

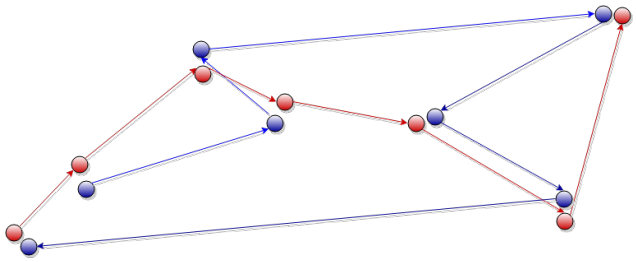


Figure 5.22 Two users taking the same buses with different order

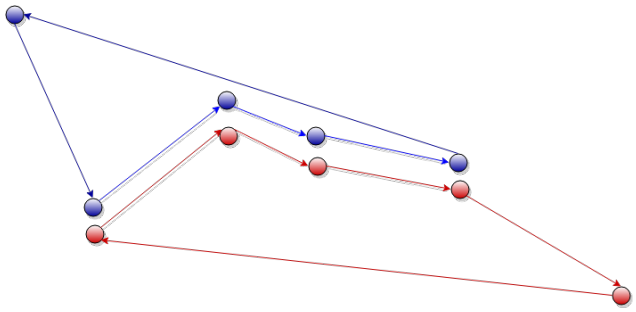


Figure 5.23 The same pattern of two users living in the different places

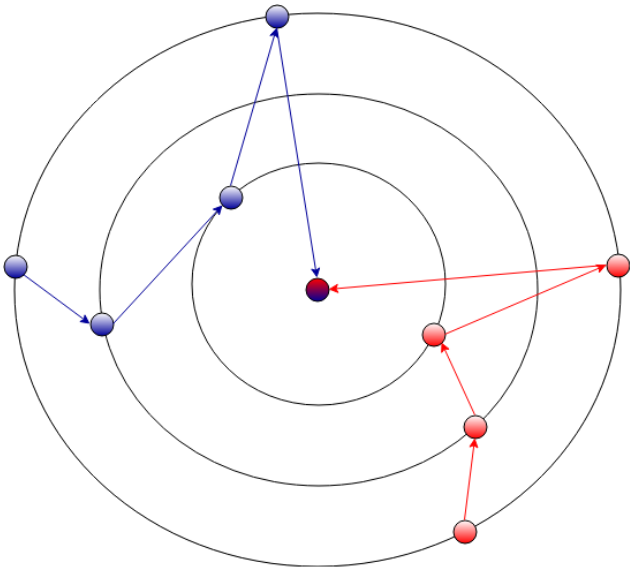


Figure 5.24 User similarity based on circular grid representation of bus stops

bus stops does not conform the uniform distribution. Despite the utilization of the bus stops usually comes from a mixture of normal distributions, for the sake of simplicity, we can assume that bus stops are sampled from just a normal distribution. Fig. 5.24, illustrates a typical public transport network, where the center of the city is the mean of the spherical normal distribution, and the off-diagonal entries of the covariance matrix are zeros, because of the spherical symmetry of the density function.

This hypothesis implies that if two bus stops are taken from the same circle with the particular radius, it can be assumed they are relatively close to each other, in contrary to the Euclidean distance. Accordingly, in Fig. 5.24, the red user is following the same pattern as the user blue (at each time point, the identical bus stops are taken from the same orbit).

So far, a number of possible use cases are introduced about the spatial public transport usages, comparing two given users. In the real world datasets, where millions of users usually take the public transport for their daily journeys, the combination of these patterns can happen in the whole picture. Moreover, taking the temporal behavior into account, certainly affects the complexity of the scheduling and methods of data analysis.

Two aforementioned questions address how the user similarity criterion can be defined under few assumptions. As the first one, we assume two users are comparable if they take the same number of bus stops in their daily trip. For the second assumption, two users are similar if in the sequence of the used bus stops each pair of the bus stops associated to the same time stamp are close to each other. Finally, by summing the all distances between pair of bus stops from an origin user, similarity of a user can be computed. One suggestion for the

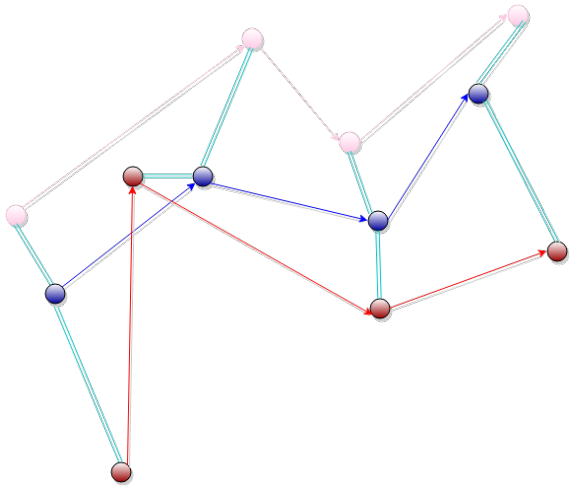


Figure 5.25 Pairwise bus stop difference criterion for measure of user similarity

origin user, is the mean geographical coordinates of the used bus stops at each time point. These few hypotheses preserve the defined constraints such that resultant traversed distance of two users is similar if they take similar bus stops at each time step. Fig. 5.25, shows three users, where the users red and orange are compared to the blue user. The sum of differences between all pairs of bus stop between blue and red circles (green lines) identifies the similarity of user blue and red. Similarly, the similarity of users blue and orange can be computed.

Formalizing this definition mathematically, suppose these two sequences are given as S_1 and S_2 from the same length. Each entry of the sequence, consists of (x, y) geographical coordinates of the bus stop. Hence, we define the similarity of two sequences as the summation of Euclidean distances of the point-wise elements. Then we have,

$$D_{\text{Euclidean}}(S_1, S_2) = \sum_{i=1}^{n} d(S_{1i}, S_{2i}) \tag{5.8}$$

where n is the number of boardings.

In addition, *Cosine* similarity and *Pearson* similarity are the other measurements suggested in (Li *et al.*, 2008) as follows,

$$D_{\text{Cosine}}(S_1, S_2) = \frac{\sum_{i=1}^{n} S_{1i} S_{2i}}{\sqrt{\sum_{i=1}^{n} S_{1i}^2} \sqrt{\sum_{i=1}^{n} S_{2i}^2}}$$

$$D_{\text{Pearson}}(S_1, S_2) = \frac{\sum_{i=1}^{n} (S_{1i} - \overline{S_1})(S_{2i} - \overline{S_2})}{\sqrt{\sum_{i=1}^{n} (S_{1i} - \overline{S_1})^2} \sqrt{\sum_{i=1}^{n} (S_{2i} - \overline{S_2})^2}}$$

5.9 Spatial-temporal data analysis with forestogram

As it was mentioned earlier, tackling the spatial analysis of the transit data is challenging due to different scenarios that may arise in defining a good metric to express the similarity among a pair of spatial trajectories. However, in analyzing the spatial-temporal data, often each component of the data, e.g. time and space is taken into account separately. One component is acting as a weight to influence the second one. Here, we take the advantage of both components reciprocally such that Euclidean property of geodesic distance for spatial data can contribute to elaborate the temporal similarity. Furthermore, sequential occurrence of time series utilization can act as a latent variable to represent the spatial structure of user behavior in public transit network. In this regard, customized version of forestogram library designed for biclustering is employed to extract the similar group of users with their corresponding temporal and spatial pattern simultaneously. To this end, daily temporal usage requires to split into hourly binary intervals as it is shown before. Then the spatial coordinate pair of GPS location (x, y) from the corresponding time interval takes the same place in the input data matrix. This way, we can perform the Euclidean distance to the entries of the matrix while the latent time information is implicitly playing its role. Now the remaining part is to modify the distance for columns of the matrix, such that (x, y) location is considered as a unit measurement for computing the dissimilarity measure associated to the Lance-William property. This way spatial-temporal patterns can be easily extracted from the columns of the matrix while similar users appear on rows. Yet hierarchical structure of the forestogram determines the evolution of the biclusters with benefits of FORIC as a statistically meaningful guide to guess how many blocks we have in the data. For better understanding what this biclustering does on a spatial-temporal matrix, let's consider an example. Suppose we have 8 users that only commute from one point to another as is shown on the map in figure 5.26 at two different timestamps. The Table in 5.2 encodes this behavior to be fed into the modified forestogram.

Therefore in order to find the spatial-temporal patterns in the given example, forestogram with adjusted distance function can easily be used here. In figure 5.27 the result of running forestogram on this example is shown where we have two main clusters of users each has two subgroups. The first one contains subgroups with two opposite directions and in the second one we have two clusters with different starting points but the same destination. Moreover, from the spatial-temporal patterns, there exists two main groups where times $\{2, 5\}$ are in one temporal behavior and $\{1, 3, 4\}$ hours belong to the same pattern. From the spatial viewpoint, we just have $(10, 50)$ location shared between two spatial behaviors because it is used in two different timestamps. This example is an illustration of the idea how forestogram

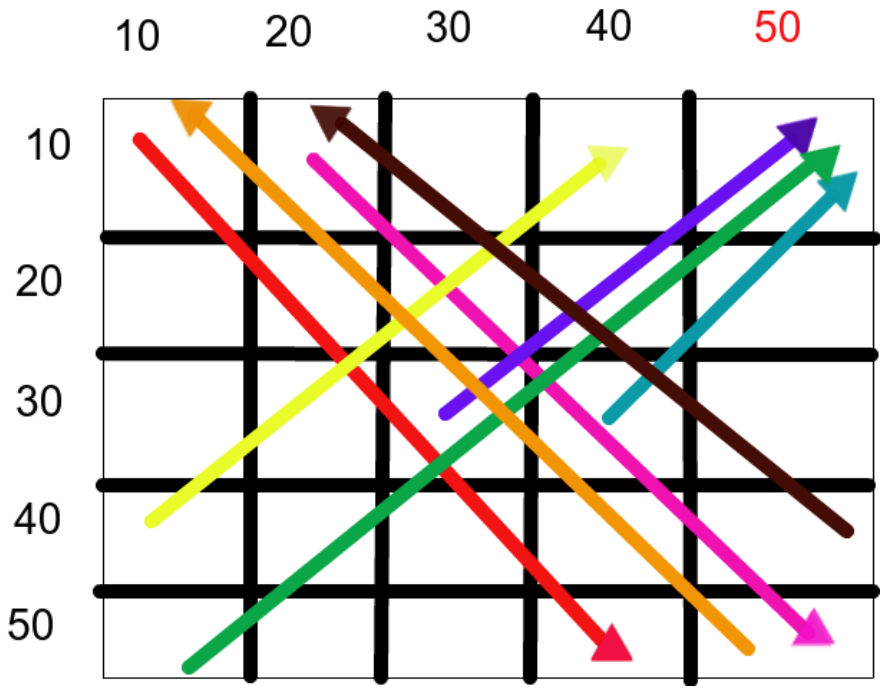


Figure 5.26 Visualization of the synthetic example of spatial-temporal data associated with 8 users and the corresponding spatial usages during 5 hours shown in Table 5.2.

can be effectively tailored for spatial-temporal data analysis such that Euclidean distance is relevant to location history and the time series is implicitly taken into account.

We also take another empirical data study besides this intuitive example to investigate this method on real data gathered for a period of one week from 113692 observations between 5 to 23 based on daily hours. It turns out forestogram in conjunction with FORIC extract 11 groups of similar users with 3 different spatial-temporal patterns described in Figure 5.29, 5.30, and Figure 5.31 with the corresponding forestogram depicted in Figure 5.28.

Three spatial-temporal patterns across 11 groups of similar users that are discovered by FORIC on the forestogram shown in Figure 5.28, are elucidated in Figure 5.29, Figure 5.30, and Figure 5.31. In both figures x axis encodes the discrete hourly usages and y axis shows the shared location in (latitude, longitude) pair. The overview of the extracted patterns that are released by forestogram, demonstrates that three temporal patterns exist in the current collected data with diversity of location histories. Here are a number of salient patterns which describe the daily behavior of subscribers in the public transit network. Late night and middle day commuters with corresponding geographical location is shown in Figure 5.29(b). Figure 5.30(b) shows the early morning and afternoon behavior of users for the related locations. In Figure 5.31(c) a combination of early morning, noon and overnight usage is depicted in three sub-blocks of bicluster-9 with variety of spatial patterns. The temporal patterns of Figure

Table 5.2 Synthetic example of spatial-temporal data associated with 8 users and the corresponding usages during 5 hours. Spatial location is denoted by (latitude, longitude) pair.

User	1	2	3	4	5
X_1	0	$(10, 10)$	0	0	$(50, 40)$
X_2	0	$(10, 20)$	0	0	$(50, 50)$
X_3	$(30, 30)$	0	0	$(10, 50)$	0
X_4	$(30, 40)$	0	0	$(10, 50)$	0
X_5	$(50, 10)$	0	0	0	$(10, 50)$
X_6	$(40, 10)$	0	0	0	$(10, 40)$
X_7	0	$(50, 50)$	0	0	$(10, 10)$
X_8	0	$(40, 50)$	0	0	$(10, 20)$

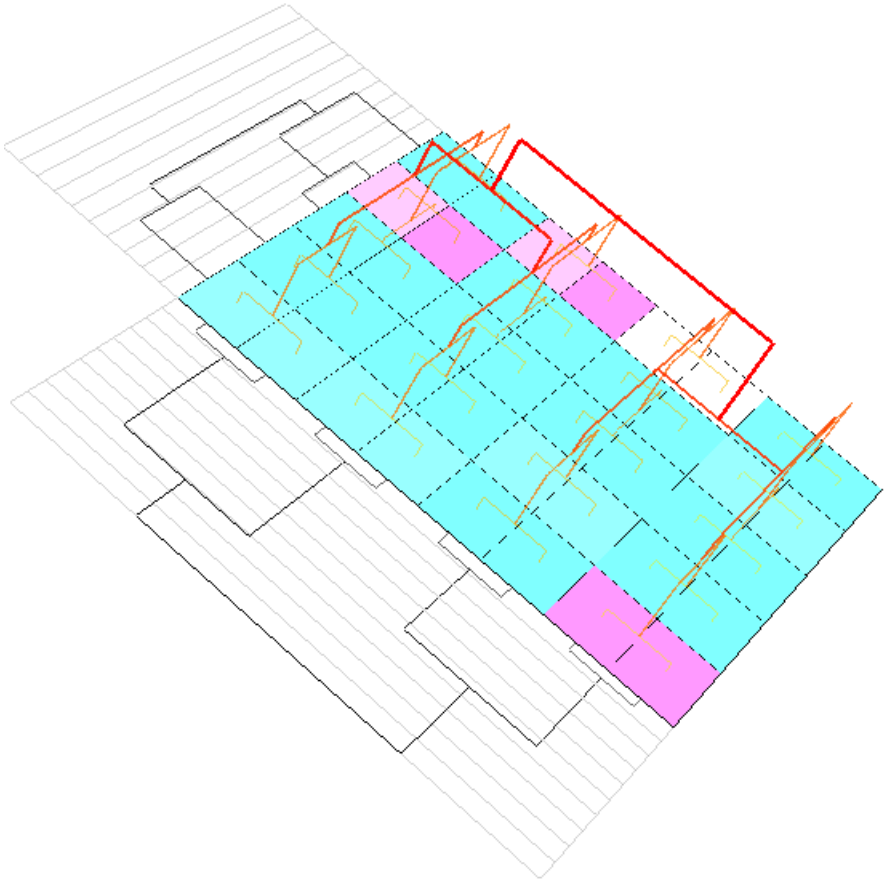


Figure 5.27 Forestogram of the synthetic example of spatial-temporal data defined in Table 5.2.

5.30(e) are very similar to the temporal patterns in Figure 5.31(c) with slightly sparse hourly usage, though with disparate geographical locations. In Figure 5.31 the second sub-block in

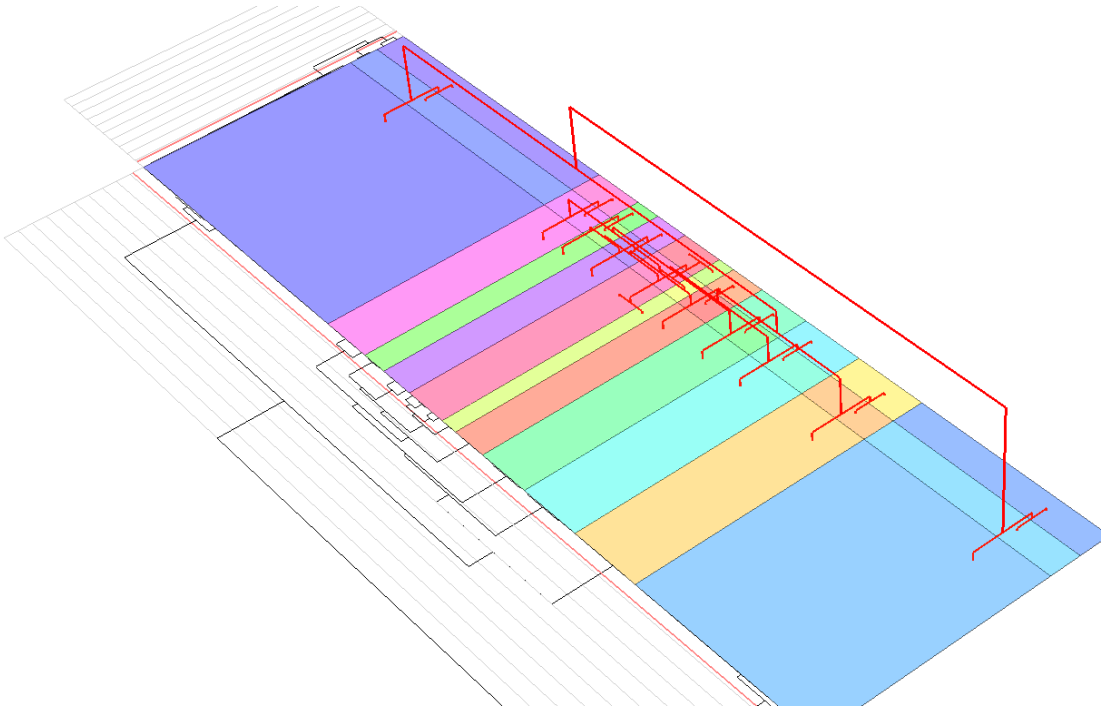


Figure 5.28 Forestogram built on top of the cluster centers obtained from the real data.

biclusters (c), and (d) shows a similar spatial patterns with the same hourly order but variant shift such that we can conclude the commuters travel from home to work and from work to home during the weekdays and weekends, respectively. The overall spatial behavior of each temporal bicluster is illustrated in Figure 5.31(f) by averaging on all users. We can argue that each bicluster identifies a certain spatial-temporal pattern in terms of scheduling, working day, weekend, summer, holidays, particular events, station/stop quality and functionality, tap in/tap out information, etc. if the type of each card was available, the location of station/stop was determined and the range of boarding date was beyond just one week.

We show the pattern of spatial-temporal behavior of the similar users extracted by forestogram in Figure 5.29, Figure 5.30, and Figure 5.31, where the temporal usage is captured on x-axis and the y-axis displays the average of the spatial data corresponding to the used hours. This new spatial-temporal patterns can help public transport analysts comprehend how the temporal pattern varies over the geographical locations inside the existing biclusters. It is worth to mention that each bicluster is also self descriptive such that, we can observe how many distinct spatial-temporal templates is discovered by the forestogram automatically to make sense of public transport data visually.

As it was discussed earlier, this subsection has not published yet and we use a preliminary limited data to conduct a pilot study for evaluating the feasibility of deploying our suggested

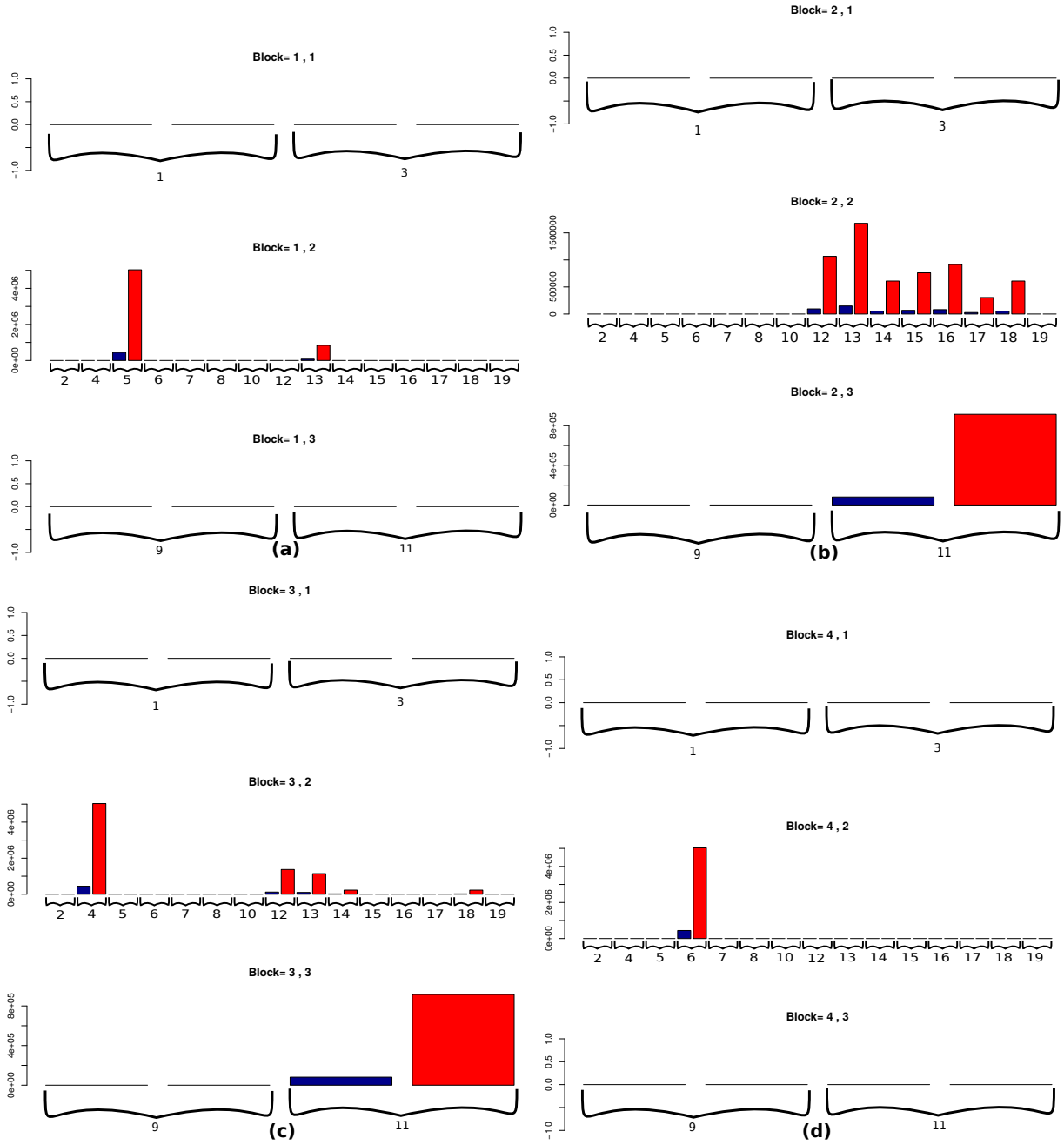


Figure 5.29 Patterns of spatial-temporal behavior extracted from the real data with modified forestogram. x axis encodes the discrete hourly usages and y axis shows the shared location in (latitude, longitude) pair.

forestogram in the area of public transport as the first application of this thesis. In light of recent development of forestogram, and promising analysis of one week data that was partially available for this experimental study, we can suggest that using the geodesic distance trick

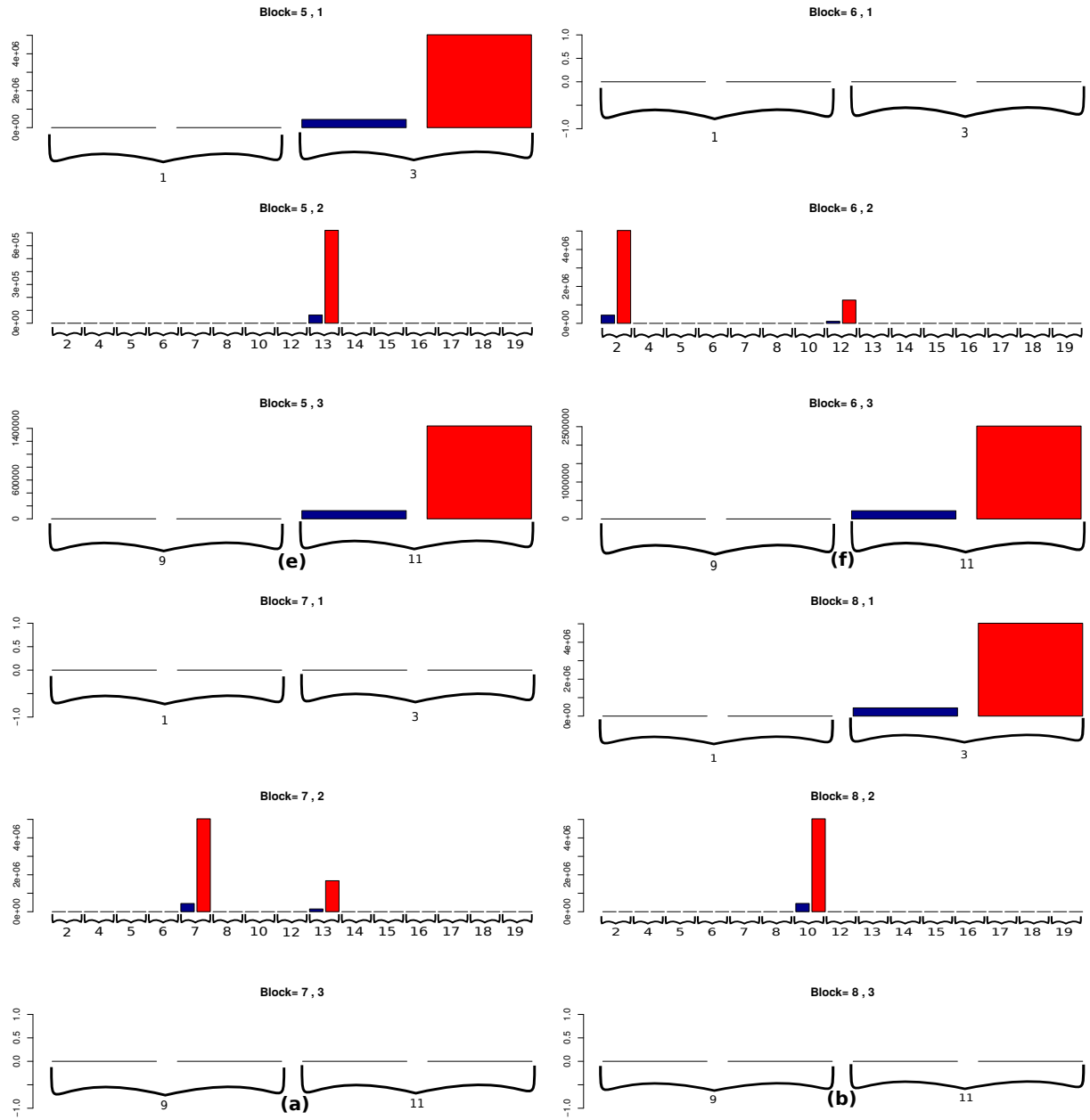


Figure 5.30 Patterns of spatial-temporal behavior extracted from the real data with modified forestogram. x axis encodes the discrete hourly usages and y axis shows the shared location in (latitude, longitude) pair.

to extract the spatial patterns through the latent temporal usage is a novel idea for spatial-temporal data analysis in public transport domain. Furthermore, the combination of SCP idea for temporal behavior along with the bicluster analysis of spatial-temporal pattern can open a new direction to study the integration of both elements from smart card data more

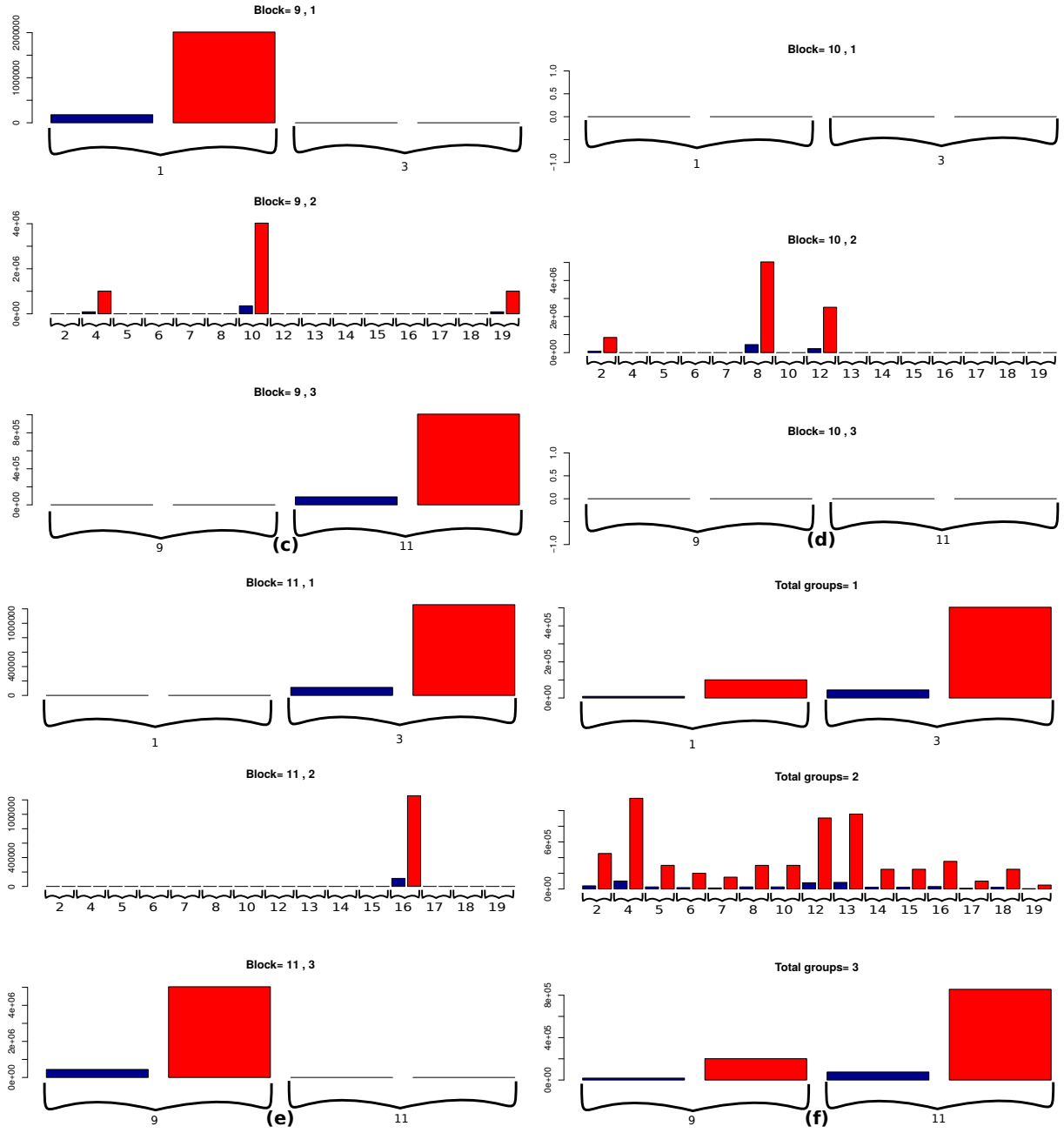


Figure 5.31 Patterns of spatial-temporal behavior extracted from the real data with modified forestogram cont. x axis encodes the discrete hourly usages and y axis shows the shared location in (latitude, longitude) pair.

effectively in the future.

CHAPTER 6 MULTIOMICS ANALYSIS OF HOST RESPONSE TO PREGNANCY

6.1 Abstract

Motivation: Despite the well-established impact of baby development during the early months of pregnancy on long-term outcomes, the biological mechanisms that govern pregnancy have not been studied in details. Most clinical assays (e.g., those based on ultrasound) can only capture abnormalities at a late pregnancy stage. The maintenance of pregnancy relies on a finely-tuned balance between tolerance to the fetal allograft and protective mechanisms against invading pathogens. This is achieved through a series of symbiotic interactions between different biological modalities. Demonstrating the chronology of these adaptations to a term pregnancy provides the framework for future studies examining deviations implicated in pregnancy-related pathologies including preterm birth and preeclampsia.

Results: We perform a multiomics analysis of 51 samples from 17 pregnant women, delivering at term. The datasets include measurements from the immunome, transcriptome, microbiome, proteome, and metabolome of samples obtained simultaneously from the same patients. Elastic net algorithm is used to measure the ability of each dataset to predict gestational age. Using stacked generalization, these datasets are combined into a single model. This model not only significantly increases the predictive power by combining all datasets, but also reveals novel interactions between different biological modalities. Future work includes expansion of the cohort to preterm-enriched populations and in vivo analysis of immune-modulating interventions based on the mechanisms identified. Furthermore, we investigate the performance of forestogram for biclustering task where selected variables can show how much supervised information contributes to the prediction of pregnancy trimesters. Additionally, visualization property of forestogram provides a comprehensive tool for analyzing the interrelated features across the pregnancy trimesters to interpret the results.

6.2 Introduction

Recent technological advances in science provide novel opportunities to unravel the complex biology of pregnancy. A particularly pressing issue is to identify the biological pathway and the converging pathological processes that lead to preterm birth (Lackritz *et al.*, 2013). Preterm birth is the major cause of neonatal death, and the second leading cause of mortality in children under the age of 5 years (Liu *et al.*, 2012).

An ongoing cohort study by the March of Dimes Prematury Research Center at Stanford University exploits recent technological advances to examine an array of biological and environmental factors associated with normal and pathological pregnancies (Stevenson *et al.*, 2013). From a biological perspective, this effort has so far produced two major lines of evidence. One line sheds light onto precisely tuned chronological changes that occur during normal pregnancy. For example, a highly multiplexed cell-based assay in whole blood revealed an "immunological clock" of human pregnancy that predicts gestational age at the time of sampling (Aghaeepour *et al.*, 2017). These findings are echoed in a longitudinal analysis of cell-free, maternal RNA (Pan *et al.*, 2016). The second line points to important pathophysiological derangements. For example, dense longitudinal sampling of the vaginal microbiome revealed community composition profiles associated with preterm birth that were validated in an independent cohort (DiGiulio *et al.*, 2015; Callahan *et al.*, 2017).

Current multiomics efforts belong to two categories generally known as multi-staged and meta-dimensional (Ritchie *et al.*, 2015). In multi-staged analysis, measurements of the same biological factors are integrated at various biological levels and using different technological platforms e.g., DNA and RNA sequencing, epigenetic analysis, and proteomics assays. Notable examples include Emilsson *et al.* (2008); Schadt *et al.* (2005); Maynard *et al.* (2008); Shabalin (2012); Shen *et al.* (2009). However, modern biological studies extend well beyond these layered measurements and include various assays such as single cell analysis, imaging, mechanical measurements using wearable sensors, and clinical phenotypes. Meta-dimensional multiomics is an emerging approach that aims at combine heterogeneous datasets to identify key factors at various biological levels, their interactions with each other, and with clinical outcomes. Some studies achieve this by simply merging all available datasets into a single matrix for joint modeling Fridley *et al.* (2012); Mankoo *et al.* (2011); Holzinger *et al.* (2013). These approaches are often susceptible to biases introduced by the differential sizes, modularities, scalings, and batch effects of the included datasets. Various kernel *e.g.,* Borgwardt *et al.* (2005), and graph *e.g.,* Kim *et al.* (2012) transformations have been proposed to address this. Depending on type of analysis that is performed against an external factor, an alternative approach is to use a mixture-of-experts methods to combine the results of independent models produced on each dataset through various algorithms ranging from voting e.g., Aghaeepour and Hoos (2013) to integration of Bayesian probabilities Zhu *et al.* (2008, 2012); Akavia *et al.* (2010).

While the analysis of a specific molecular data set is of undisputed value, the meta-dimensional analysis of various data sets holds significant promise. Physiological changes during pregnancy are highly dynamic and involve multiple interconnected biological systems. The simultaneous interrogation of these systems with suitable technologies can reveal

otherwise unrecognized crosstalk. Understanding such crosstalk can inform several lines of investigation. From a biological perspective, it points to important disease mechanisms such as immune programming by the microbiome, or specific interactions between proteins and cellular elements (Aghaeepour *et al.*, 2017; Dethlefsen *et al.*, 2007). From a diagnostic perspective, it reveals biomarkers from several biological domains that provide higher predictive power if combined. Alternatively, it also points to substitute biomarkers in an accessible biological compartment, which can replace biomarkers that are difficult to obtain or expensive to measure.

The first objective of this study is to test whether a multiomics analysis of transcriptomic, immunological, microbiome, and proteomic data can increase the power of a model predicting gestational age in term pregnancy. The second objective is to probe whether and to what extent each data set contributes to the model. The third objective is to test whether the number of model parameters can be reduced without compromising predictive power. The fourth objective is to interrogate derived model for novel and testable biological links as is shown in Figure 6.1. And in the fifth objective, regardless of the supervised information forestogram is performed to demonstrate the role of biclustering in analyzing the integrative model for multiomics dataset in contrast to the other unsupervised perspectives shown in Figure 6.14.

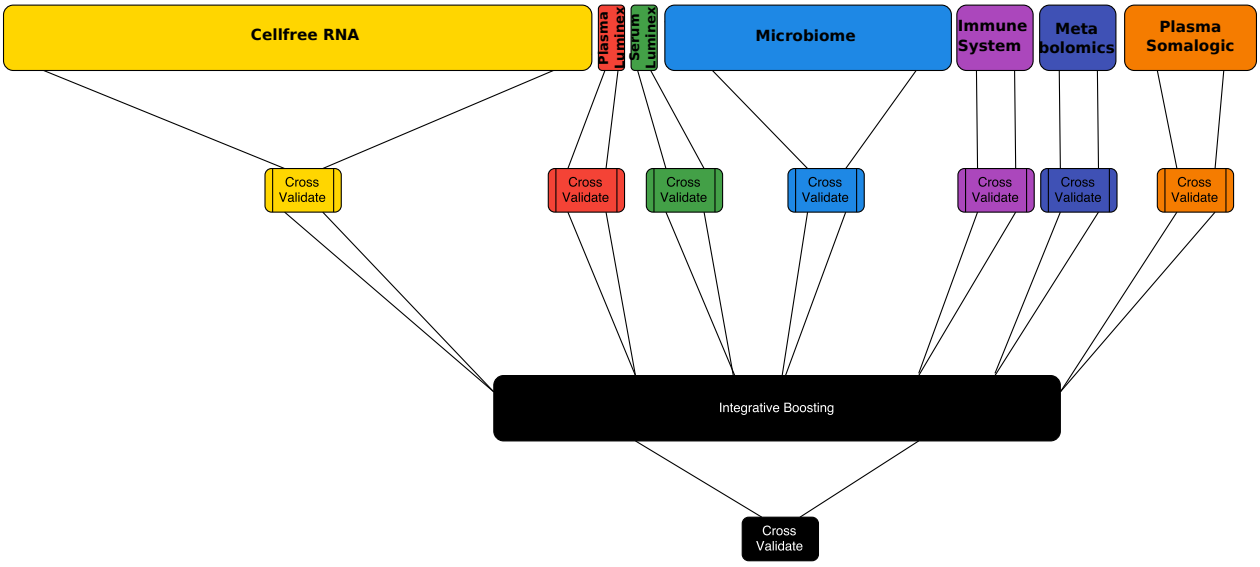


Figure 6.1 Integrative model for combining seven multiomics dataset through cross-validation. In the first layer, for each omic dataset a regression model is tuned. Then the integrative prediction is made by bringing gestational output from each omic dataset together in the second layer.

6.3 Results

6.3.1 Overview

Samples from a total of 51 visits throughout pregnancy and 17 visits 6 weeks postparturm are collected. Samples are analyzed for seven biological modalities, Cell-free transcriptomics, luminex proteomics in plasma and serum, microbiome analysis from several body sites, mass cytometry analysis of whole blood, and metabolomics and proteomics analysis of plasma Figure 6.2(a). Not only these datasets significantly varied in the number of measurements Figure 6.2(b), but also has different levels of complexity as measured by the number of principal components needed for accounting for 90% variance of each dataset Figure 6.2(c).

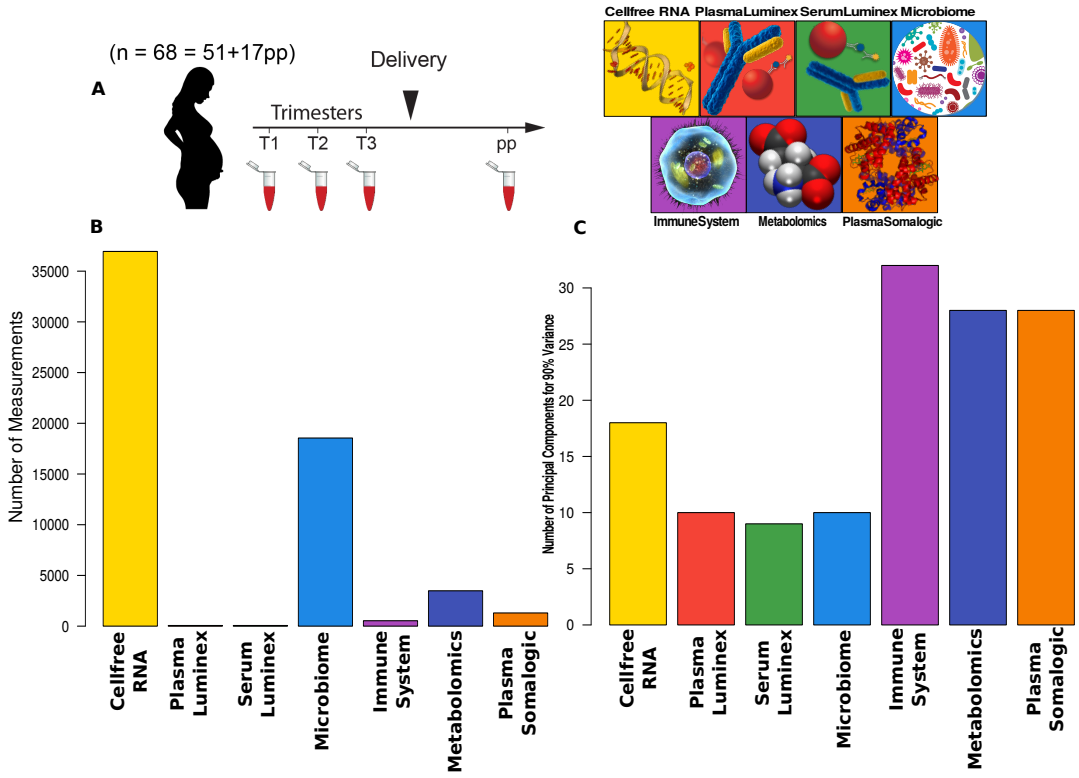


Figure 6.2 (a) Overview of the study design. A total of 51 samples are collected during three trimesters of pregnancy as well as an addition 17 samples 6 weeks after delivery. Seven datasets are produced for each sample. (b) The number of biological measurements in each dataset. (c) Complexity of each dataset calculated as the number of principle components needed to capture 90% variance.

The first step toward gestational age prediction, is to analyze each dataset separately to use the prediction of each omic prediction in the second step for the integrative model. Elastic net (Zou and Hastie, 2005) is deployed as a promising regression technique for high-dimensional small sample size dataset. In the second step, stacked generalization is designed based on elastic net to improve the prediction through the incorporation of multiomics dataset.

6.3.2 Estimation of Gestational Age

Elastic net algorithm is used to predict the gestational age of each subject at each visit. A two-layer cross-validation procedure is used to both optimize the free parameters of the elastic net model and to ensure predictions are always made on samples that are not used

for training, to avoid overfitting see Figure 6.3(a). A broad range of p-values with Plasma Proteomics analysis using the Somalogic platform producing the highest correlation 6.3(b). Results remained generally consistent on the test set 6.3(c). These findings are independent of the size or complexity of each dataset 6.2(b) and (c).

6.3.3 Stacked Generalization

The estimations produced by these models is then merged using an additional elastic net model. Cross-validation is synchronized across all layers to ensure predictions are made on samples that have not been seen by the stacked generalization elastic net or any of the models built on individual dataset Figure 6.4(a). For visualization purposes, the top hits from each model are extracted and visualized using a Minimum Spanning Tree (MST) between the selected features.

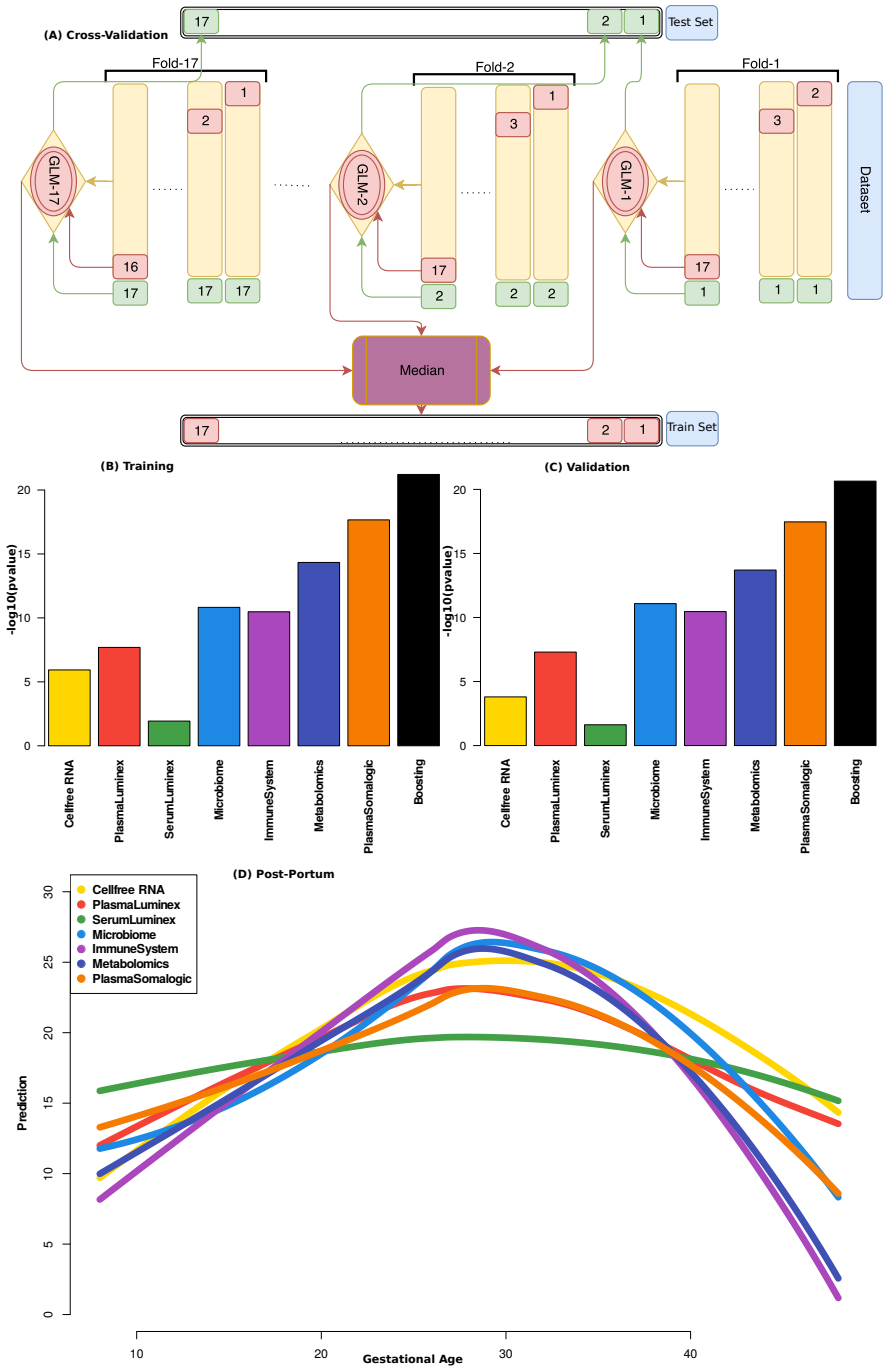
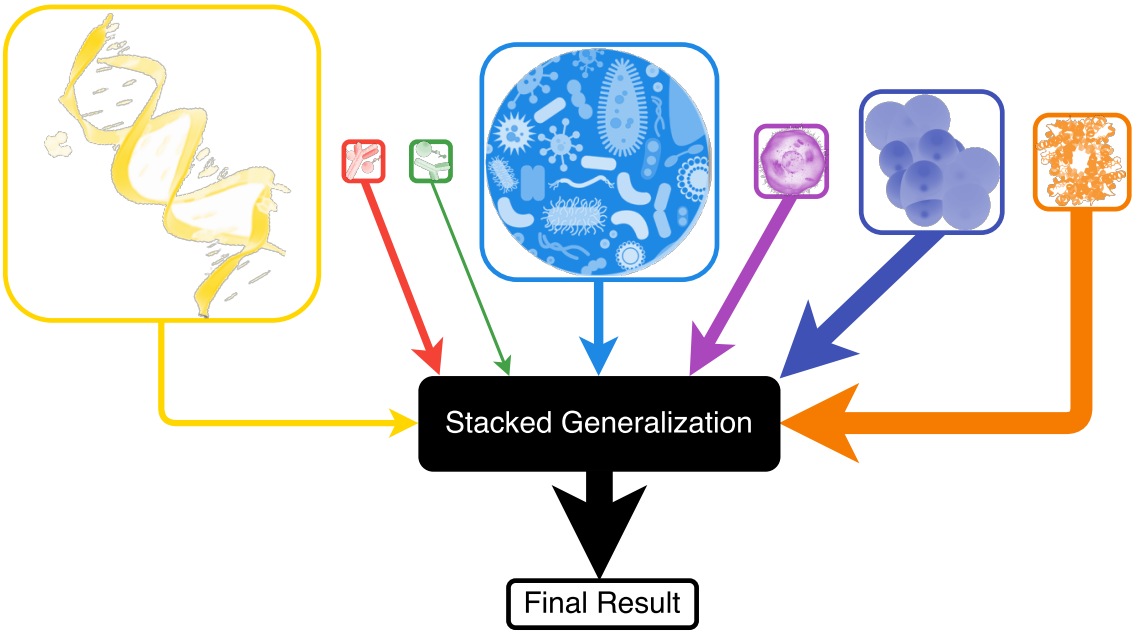
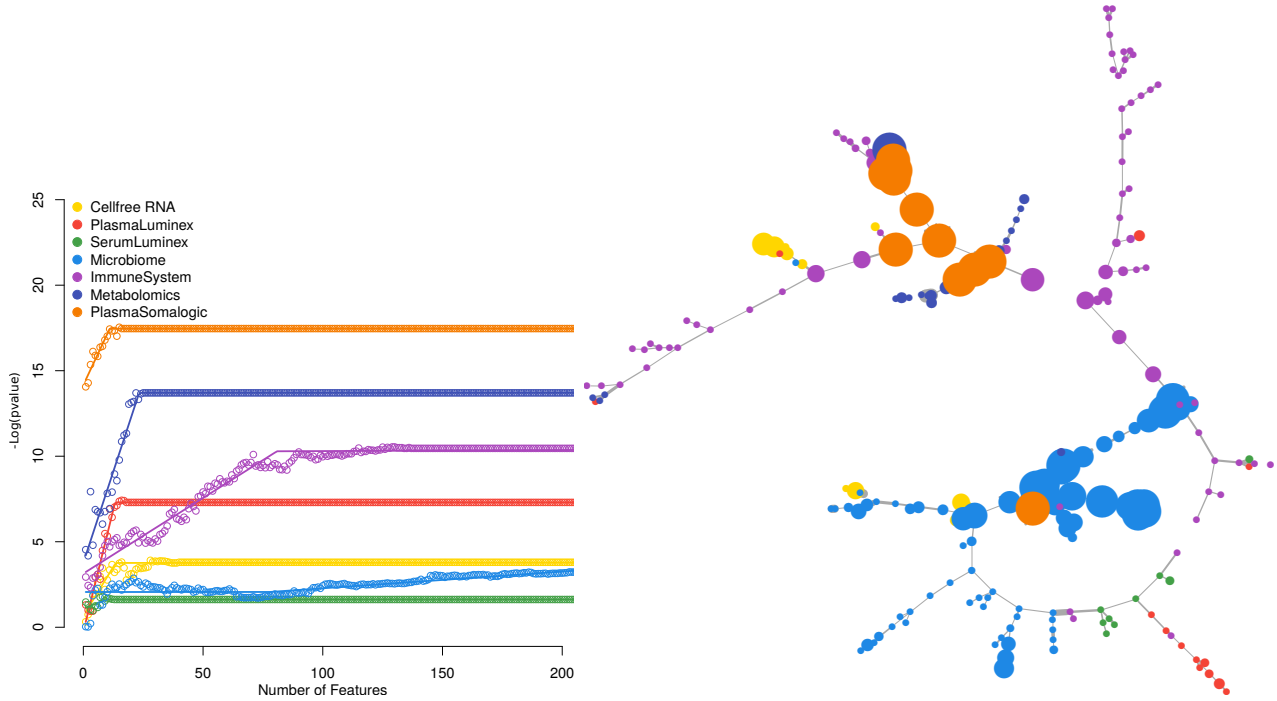


Figure 6.3 a) Overview of the two-layer cross-validation procedure. On the outer layer, a modified leave-one-patient-out cross-validation procedure is used in which all samples from the same subject (as opposed to just one subject) is left out as a blinded sample. Within each fold a second cross-validation is performed to optimize the free parameters of elastic net. (b and c) the Spearman correlation between the (b) training set and (c) test set cross-validated results for each dataset. (d) performance of the trained models on the whole datasets including the first trimesters of pregnancy and post-partum that is never exposed to the training set.

This resulted in a set of 226 interrelated features, revealing statistically robust interactions within and between each omics dataset. A Minimum Spanning Tree (MST) representation organized these interactions into a branched structure in which the distance between two features is proportional to the strength of the correlation between them. Metabolomics, transcriptomics and proteomics features primarily segregated into three clusters. Cytomic features from the immune system were distributed across the MST graph, forming a link between other omics datasets rather than being confined to a single cluster. The MST graph highlighted the connectivity between biological processes measured in the plasma (metabolomics, transcriptomics, proteomics measurement) or local compartments (microbiome data) and cell-specific immune responses measured in the peripheral blood compartment.



(a) Stacked Generalization



(b) Model reduction and MST correlation visualization

Figure 6.4 (a) Stacked generalization analysis. The size of the boxes is proportional to the log 10 of the number of measurements in each dataset. The thickness of the arrow is proportional to the $-\log_{10}$ of p-value of a correlation test for gestational age; (b) Visualization of the most predictive features in a correlation network. The size of each node is proportional to the univariate correlation between that feature and gestational age. Color represents the corresponding dataset.

With respect to the microbiome data, a strong correlation is observed between changes in the composition of bacterial species localized in the oral cavity and the frequency of B-cells and TCRgd+ T-cell, a finding consistent with the unique role of TCRgd+ T-cell in mucosal immunity. With respect to the metabolic dataset, the model reveals strong correlations between the plasma factor pregnanolone and the NF-kB signaling in myeloid dendritic cells (mDCs) and regulatory T-cell (Tregs). Pregnanolone, or $3\alpha, 5\beta$-tetrahydroprogesterone($3\alpha, 5\beta$-THP), is an endogenous steroid biosynthesized from progesterone. Modulation of immune cell function by progesterone and its derivative is well established (Druckmann and Druckmann, 2005). However, their role in regulating the function of specific immune cell subsets during pregnancy is not fully understood. The results thus generate a novel hypothesis that pregnanolone may regulate important aspects of mDC and Treg function during pregnancy.

With respect to the proteomic dataset, a three-way interaction between the transcriptomic, proteomic and cytomic datasets was particularly interesting, as it highlighted a novel connection between previously reported models of molecular clocks of pregnancy. This interaction contained the Chorionic Somatomammotropin Hormone-1 (CSH-1), represented at the transcript (cell-free RNA dataset) and protein (Somalogic dataset) levels, and the endogenous activity of the transcription factor STAT5 measured at the single-cell level in CD4+ and CD8+ T cell subsets. CSH-1 is known to bind to the prolactin receptor (Walsh and Kossiakoff, 2006), which signals through the JAK2/STAT5 signaling pathway (Gouilleux *et al.*, 1994).

The strong correlation observed between CSH-1 RNA and protein levels, and STAT5 activity in T cells (R=0.5936333, p=4.402×10^{-06}) prompted further examination in an in vitro model to determine whether CSH-1 can directly activate the JAK2/STAT5 signaling pathway in T cells. However, incubation of whole blood samples from non-pregnant or pregnant (Supplemental Figure S3) women with CSH-1 did not induce the phosphorylation of STAT5 in CD4+ or CD8+ T cell subsets. On further inspection of the proteomic dataset, CSH-1 was found to belong to a community of tightly correlated plasma factors known to regulate the JAK/STAT signaling pathway. This community included the inflammatory cytokine Interleukin-2. Supplementary Figure S3 shows that, in contrast to CSH-1 or prolactin, incubation of whole blood samples with IL-2 induced a robust STAT5 phosphorylation signal in all major T cell subsets. These results suggest that in the context of pregnancy, the progressive increase in intracellular STAT5 activity in T cell subsets is likely driven by changes in IL-2 rather than CSH-1.

6.4 Methods

Pregnant women presenting to the obstetrics clinics of the Lucile Packard Children's Hospital at Stanford University for prenatal care were invited to participate in a cohort study to prospectively examine environmental and biological factors associated with normal and pathological pregnancies. Women were eligible if they were at least 18 years of age and in their first trimester of singleton pregnancy. Samples were obtained during the first ($7-14$ weeks), second ($15-20$ weeks), and third ($24-32$ weeks) trimesters of pregnancy, and 6 weeks post-partum. In a subsets of 17, specimens were collected for the comprehensive analysis of cell-based immunological changes in whole blood using mass cytometry, proteomic changes in plasma using multiplexed antibody and aptamer-based platforms, transcriptomic changes in plasma using cell-free RNA, and microbial changes in the vagina using high resolution sequencing techniques. The study was approved by the Institutional Review Board of Stanford University School of Medicine and all participants provided written informed consent.

6.4.1 Elastic net

For a matrix $\mathbf{X}$ of all features from a given dataset, and a vector of estimated gestational ages at time of each sampling $\mathbf{Y}$, the EN algorithm calculates coefficients $\boldsymbol{\beta}$ to minimize the error term $L(\boldsymbol{\beta}) = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$. An L_1 regularization (**?**) to increase model sparsity (which facilitates biological interpretation and validation). However, this approach is not ideal for the analysis of the highly interrelated biological data sets, because it would select only representatives of communities of highly correlated features while disregarding highly correlated but potentially biologically relevant features. This limitation is addressed by using an additional L_2 regularization penalty: $L(\alpha, \lambda, \boldsymbol{\beta}) = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2 + \lambda\Big[(1-\alpha)||\boldsymbol{\beta}||_2 + \alpha||\boldsymbol{\beta}||_1\Big]$, where $||\boldsymbol{\beta}||_2 = \boldsymbol{\beta}^\top\boldsymbol{\beta}$ and $||\boldsymbol{\beta}||_1 = \sum_{i=1}^{n}|\boldsymbol{\beta}_i|$. The subset selecting factor λ controls the sparsity of the model and the smoothing factor α controls the smoothing of selection from correlated variables (Zou and Hastie, 2005).

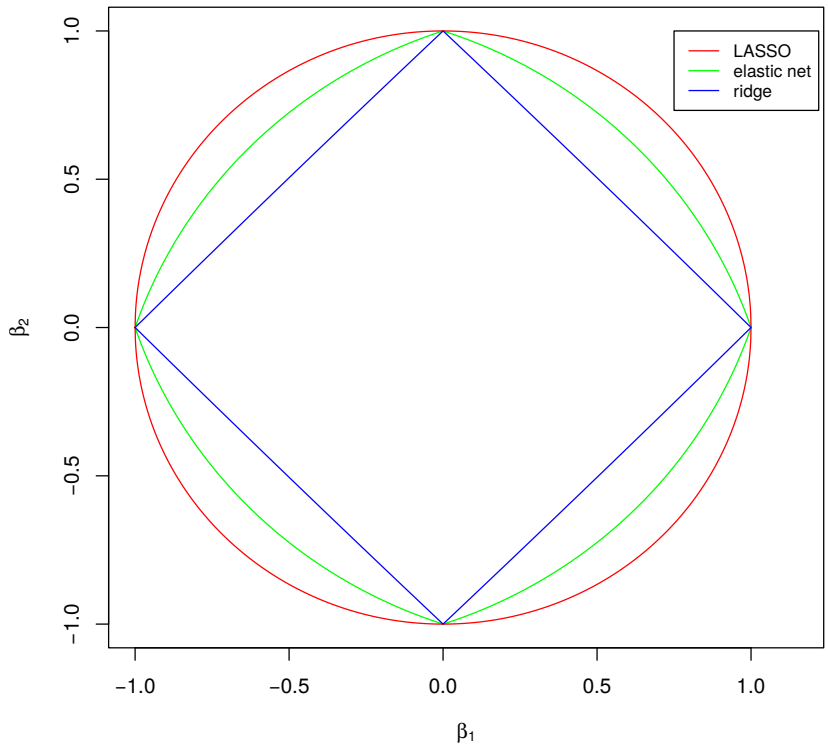


Figure 6.5 An example of bivariate elastic net penalty with $\alpha = .5$, in presence of LASSO and ridge regression constraints.

The ratio of samples to features is the key factor that affects the statistical model performance. A simple linear regression model is prone to overfit the data by adversely increasing the model complexity through dramatic number of features. Providing a model with a good generalization capacity through the overwhelming features (dimensions) with respect to the fixed number of samples requires sparse model selection techniques to avoid the curse of dimensionality. Although *LASSO* (Tibshirani, 1996) with ℓ_1 regularization penalty has been shown to be an effective sparsification model by setting irrelevant variables to zero, certain drawbacks arise due to growth of highly correlated features. Since, in biology often relatively small group of correlated measures among many other features are associated to particular disease, elastic net incorporates the ℓ_2 constraint with ℓ_1 penalty to select the correlated group of features for accounting the same biological pathway see Figure 6.5.

$$L(\alpha, \lambda, \boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2 + \lambda\Big[(1-\alpha)||\boldsymbol{\beta}||_2 + \alpha||\boldsymbol{\beta}||_1\Big]$$

6.4.2 Cross-validation

Model selection in elastic net for the hyper-parameters, smoothing factor α and sparsity penalty λ is performed with two-layer leave-one-patient-out cross-validation to avoid over-fitting the training data. In this setting, the patients that are left out in the first layer are constituting the test set, while the patients of the second layer create the training data for hyper parameters tuning. In this regard, one patient sampled at three trimesters of pregnancy is left out as unseen data for reporting the test p-values in the first layer of cross-validation. Next in the second layer, another patient is left out for model selection and computing the train p-values. In the first layer of cross-validation, we repeat this procedure 17 times to keep every single patient out once in the iteration while in the second layer a similar routine carries out 16 times for the remaining patient. Thus, in the learning phase no patient is seen at all by the elastic net except for calculating the training p-values. Then, the optimized hyper-parameters α and λ are selected to output the test p-values as is demonstrated in Figure 6.3.

6.4.3 Stack generalization

Ensemble learning is categorized into two different types, 1) diverse models on the same data, 2) same model selection technique on diverse data (Sharkey, 1996). Various methods are proposed for diversification of the algorithms to perform on the same data. Alternating the data by processing, sampling or bootstrapping is a another hybrid approach that falls in between the two categories (Sharkey, 1996). Here, we emphasize on applying the fixed model selection topology on multiple sources of data where an effective algorithm is needed to combine the output of several models. Averaging and its weighted variant is a common choice for linear pooling aggregation. In contrary, nonlinear perspective is also investigated by voting, rank-based algorithm, and order statistics. Dempster-Shafer is another nonlinear approach for fusion of information under uncertainty with several alternatives including Bayesian networks, fuzzy logic, neural networks and probability theory. The next popular nonlinear approach is stacked generalization, where the outputs predicted from the feature space is given as the input to next level of features for prediction (Sharkey, 1996). Additionally, stacked generalization, is designed for minimizing the generalization error by decreasing the bias with many applications in biology (Wolpert, 1992; Breiman, 1996; Wang *et al.*, 2006; Ge and Wong, 2008; Larranaga *et al.*, 2006; He *et al.*, 2013; Yang *et al.*, 2010).

To combine several multiomics dataset, there are two level of integration in 1) feature space 2) output of prediction space. For the first level of abstraction, merging all features to reconstruct a whole new dataset can be considered to combine all seven multiomics dataset.

In the output level, stacking the predictions corresponding to each dataset, builds up an abstraction representation for the multiomics dataset in the reduced space. To this end, *stacked generalization* ensemble is deployed to combine the results of multiple elastic nets learned from the multiomics dataset via synchronized leave-one-patient-out cross-validation. Figure 6.9, and Figure 6.8 shows the prediction of each dataset with respect to the gestational age, and also the stacked generalization model.

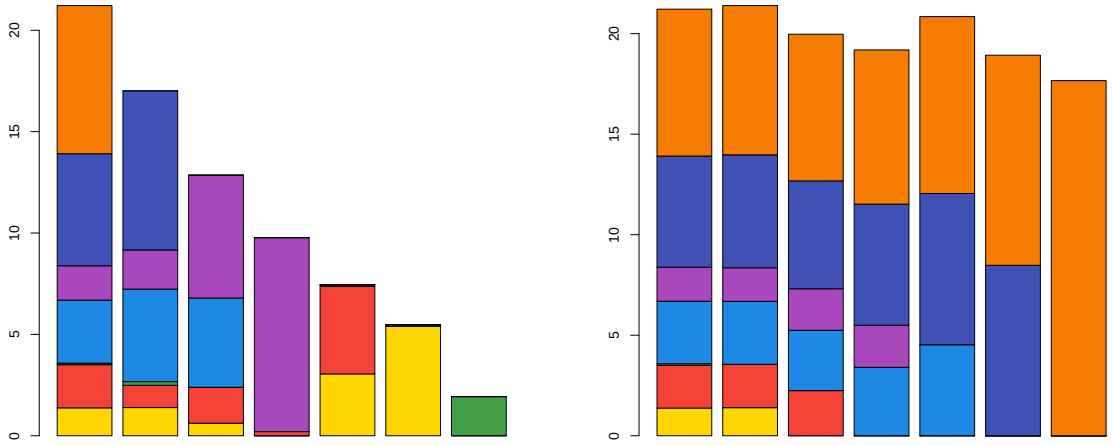


Figure 6.6 Ablation (left) and inverse ablation (right) analysis of each dataset's contribution in the integrative model. Elimination of each dataset is carried out according to the p-value of gestational age prediction shown in Figure 6.4 in ascending, and descending order, respectively. Color portion is associated with the coefficient of each dataset represented by the stacked generalization integrative model.

In particular circumstances due to the limited access to all seven technologies especially deprived areas, only a subset of multiomics dataset is available. In this regard, it is vital to determine the most important omics measurements to make the prediction. Figure 6.6 shows the effect of each omic dataset toward integrating the multiomics dataset for gestational age prediction. Ablation and inverse ablation analysis of each dataset's contribution in the integrative model. Elimination of each dataset is carried out according to the p-value of gestational age prediction shown in Figure 6.4 in ascending, and descending order, respectively. Color portion is associated with the coefficient of each dataset represented by the stacked generalization integrative model. In addition to the elastic net, other state-of-the-art regression techniques are tested as is shown in Figure 6.10. The hyper parameters of each

method are tuned by the two-layer leave-one-patient-out cross-validation procedure for predicting the gestational age on the test set. Elastic net predominantly outperforms the other rival methods especially for the integrative model.

6.4.4 Correlation network

Interrelated features extracted from different multiomics dataset are combined together in the context of correlation network such that the edges reflect the adjusted correlation among the multiomics features. The node's size represents the magnitude of the corresponding elastic net coefficient. The group of features selected from the same dataset is differentiated by the associated color shown in barplots in Figure 6.3(c). Moreover, in Figure 6.7, the visualization network is shown where the correlation direction is denoted by the intensity of blue and red colors indicating the negative or positive correlation, respectively. All p-values are adjusted using Bonferroni's method (adjusted-p-value=$\frac{p-value}{n}$), where n is the number of features.

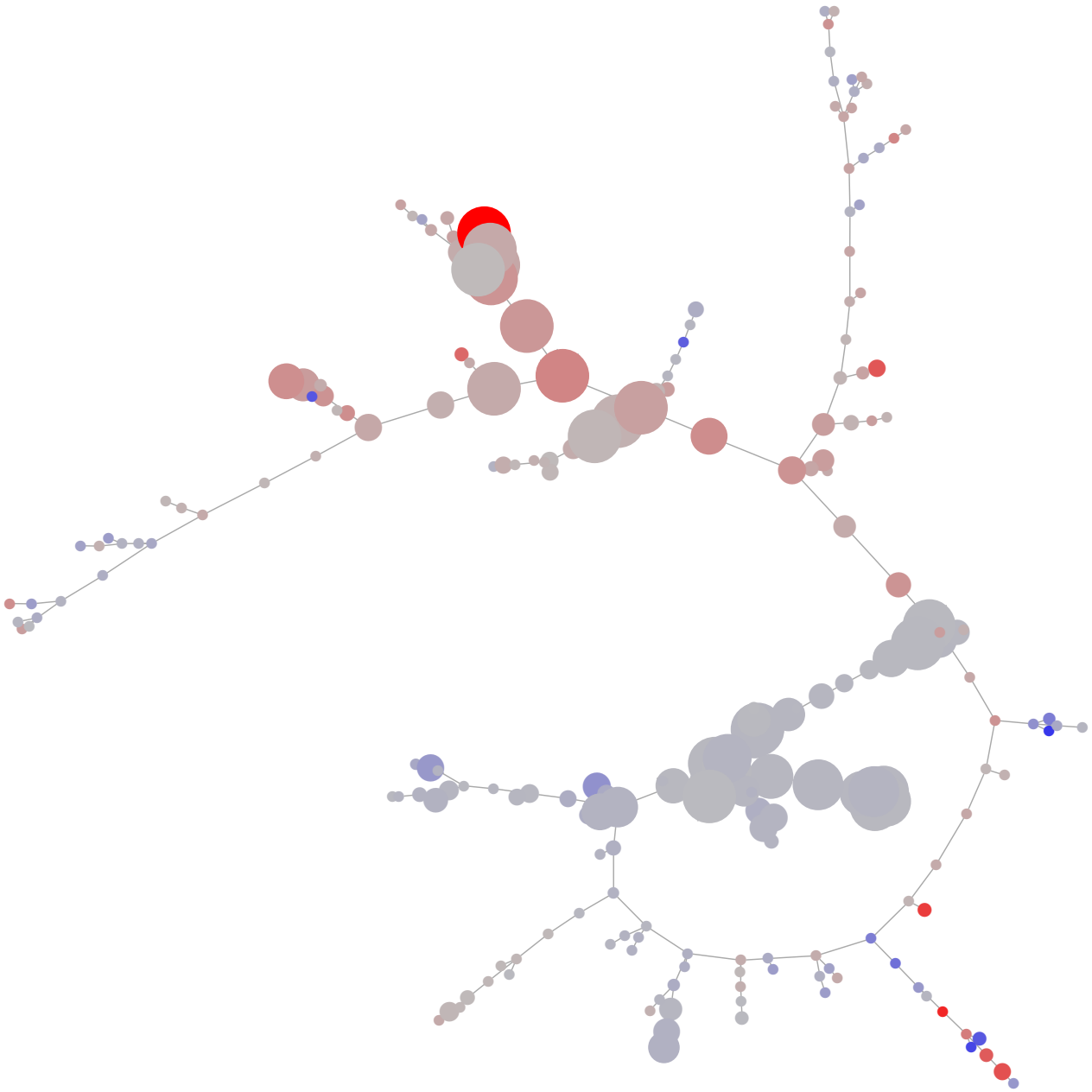


Figure 6.7 Correlation network of interrelated features extracted from different multiomics dataset. An edge reflects the adjusted correlation among the multiomics features. A node's size represents the magnitude of the corresponding elastic net coefficient. Correlation direction is denoted by the intensity of blue and red colors indicating the negative or positive correlation, respectively.

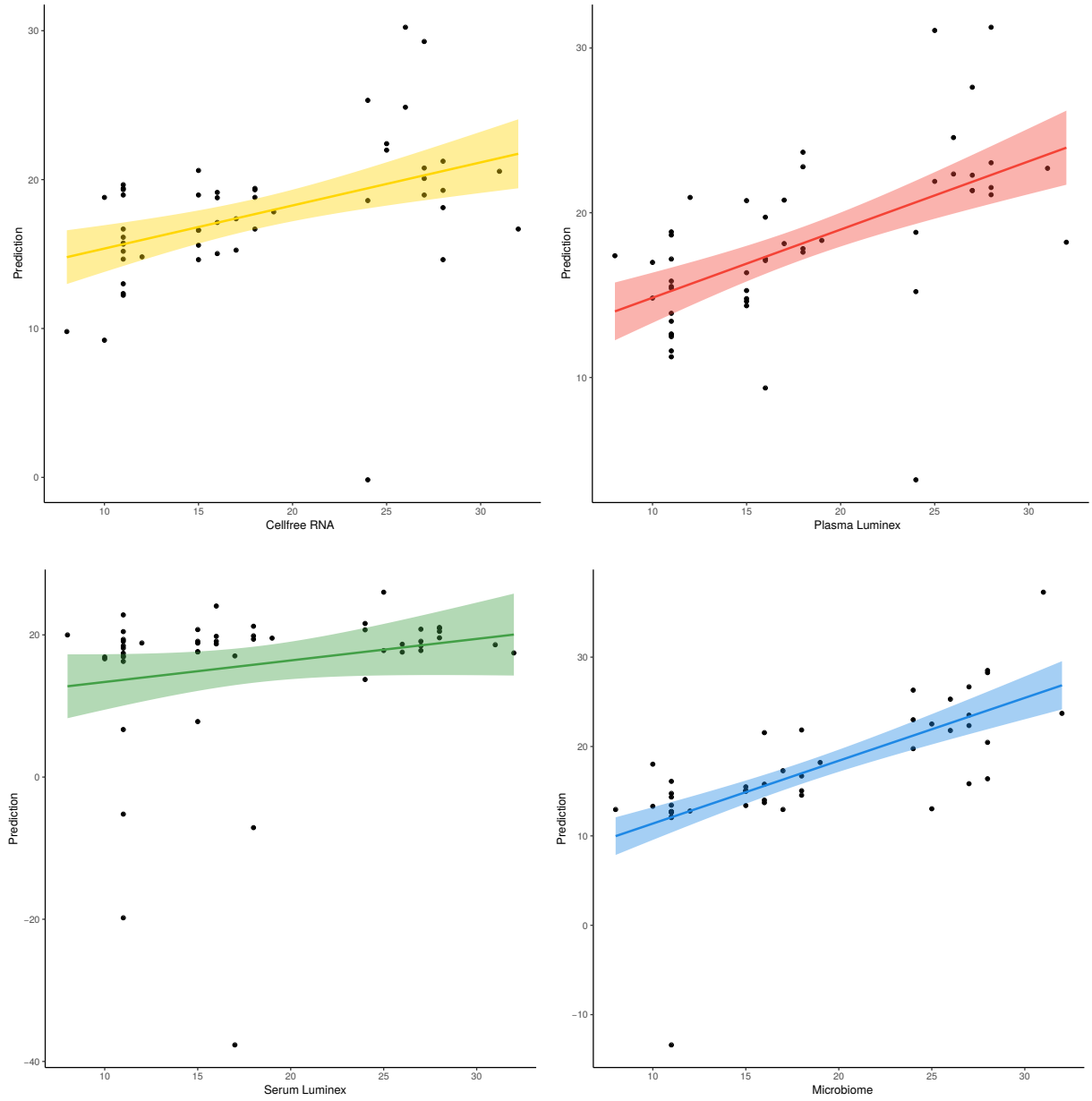


Figure 6.8 Regression lines between actual gestational age and the corresponding predictions from seven multiomics dataset and stacked generalization with their 95% confidence interval.

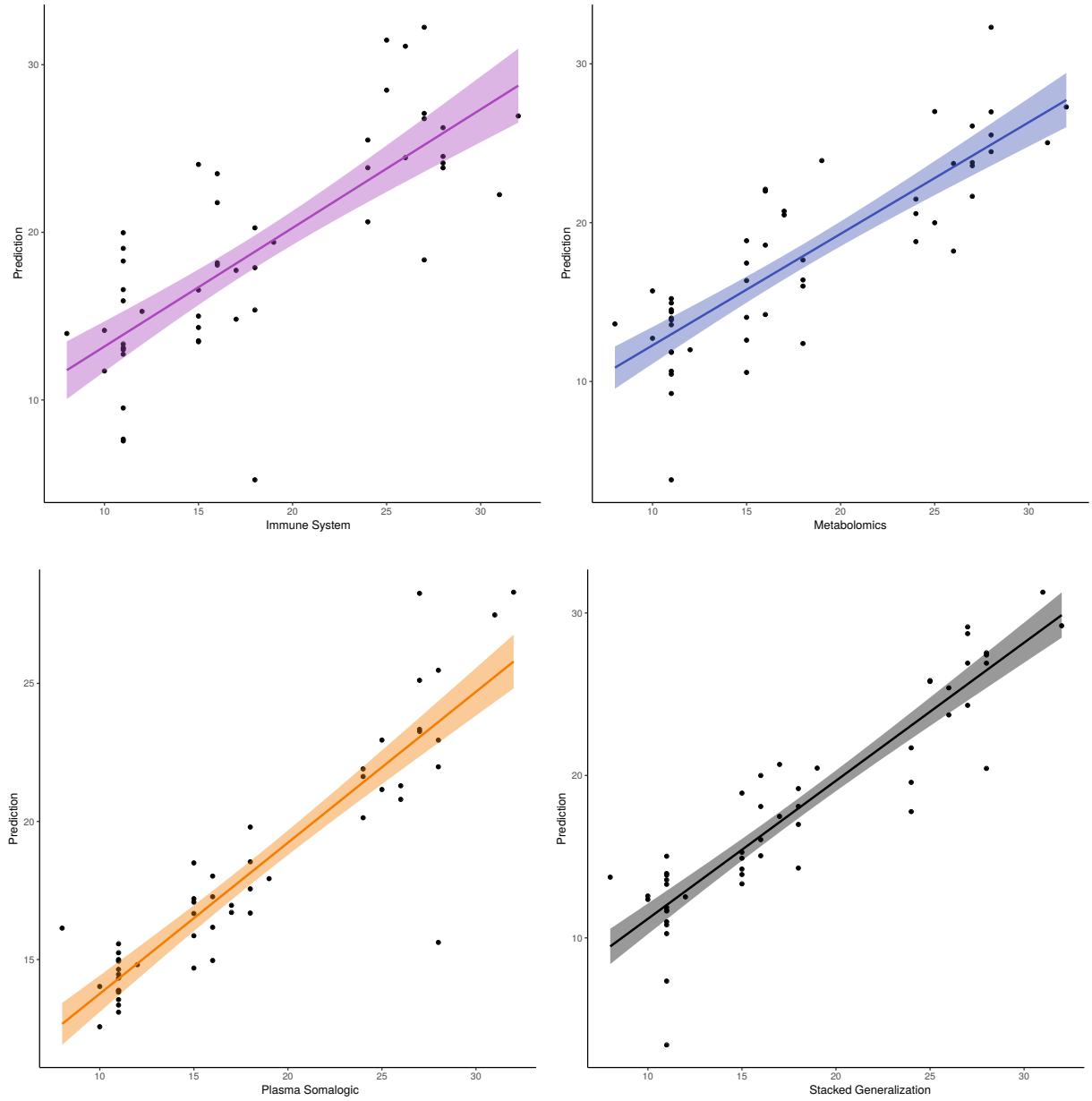


Figure 6.9 Regression lines between actual gestational age and the corresponding predictions from seven multiomics dataset and stacked generalization with their 95% confidence interval cont.

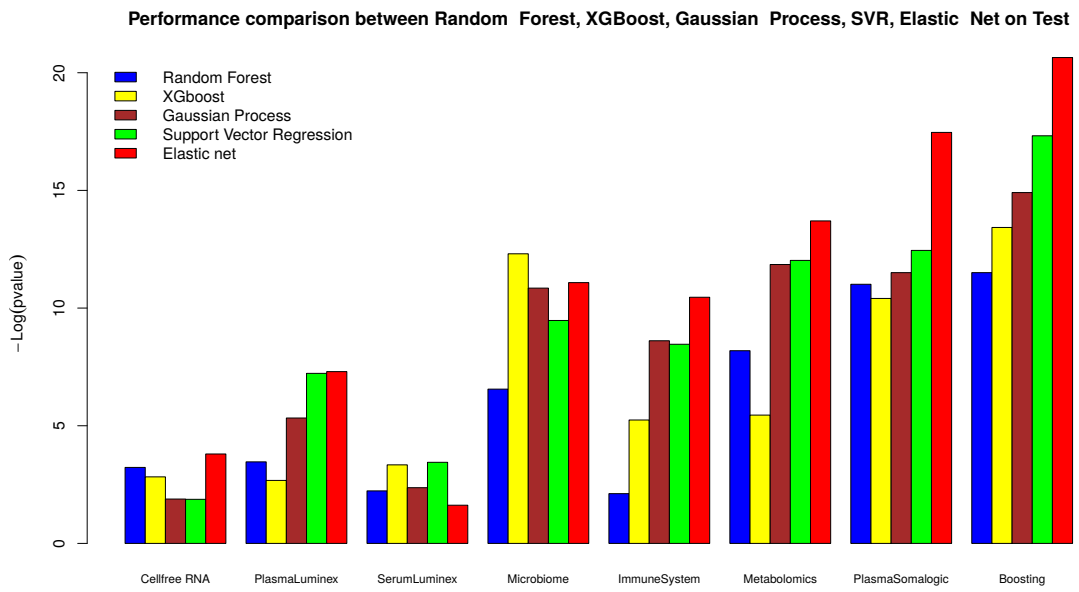
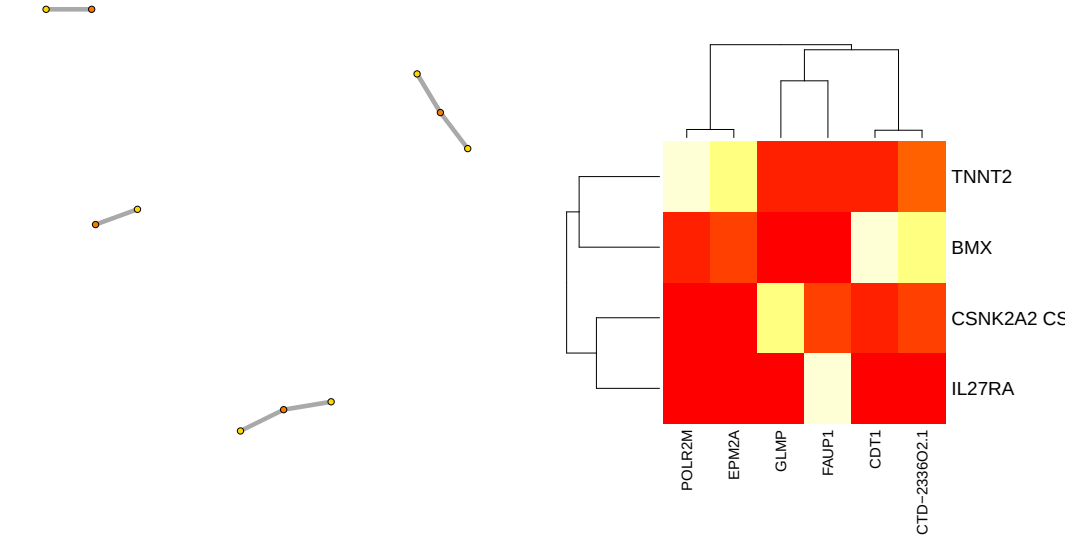


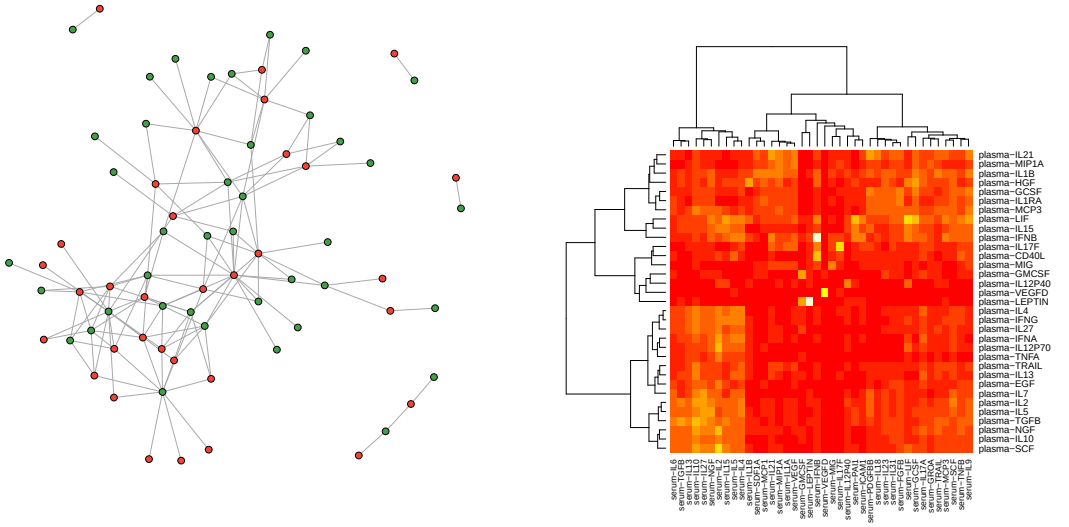
Figure 6.10 Overview of performance comparison using a number of regression algorithms, e.g. random forest, XGboost, Gaussian process, support vector regression, and elastic net. The hyper parameters of each method are tuned by the two-layer leave-one-patient-out cross-validation procedure for predicting the gestational age on the test set. Elastic net predominantly outperforms the other rival methods especially for the integrative model.

6.5 Unsupervised Analysis

Supervised integrative model based on elastic net for predicting the gestational age is developed and discussed across seven omics data. However, the available data of this project can be used for other biological and clinical purposes as well. In this section we propose to investigate an unsupervised integrative model with the introduced forestogram framework to combine all seven datasets. Before biclustering analysis, we analyze the data with two other unsupervised methods to figure out how datasets are intercorrelated with each other. Spearman's rank correlation coefficient is one of the unsupervised approaches that is known as nonparametric statistic (Spearman, 1904). In contrary to the Pearson's correlation, this method is a nonparametric measure of a monotonic relationship. Moreover, the calculation of Spearman correlation coefficient is based on the data rank rather than the data continuity. In Figure 6.11, and Figure 6.12 the Spearman's rank correlation coefficient for a few pair of datasets are shown in network visualization and heatmap representation.

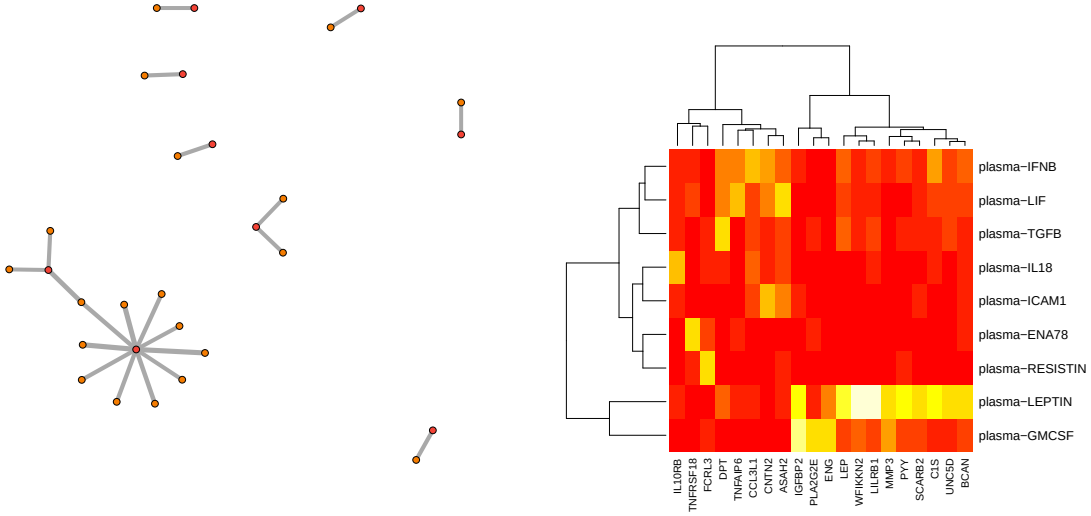


(a) Rank correlation across Cell-free RNA and PlasmaSomalogic

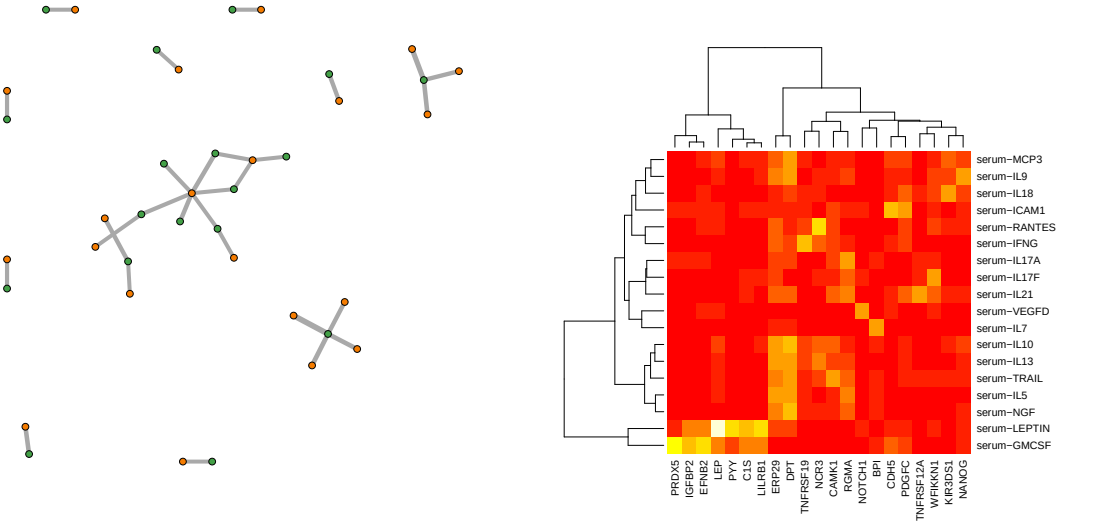


(b) Rank correlation across PlasmaLuminex and SerumLuminex

Figure 6.11 Illustration of rank correlation among a number of datasets. Left panel shows the network representation of RGCCA after Bonferroni adjustment such that presence of an edge between a pair of nodes shows strong correlation between those nodes. Right panel simply demonstrates the heatmap visualization of correlation among two datasets.



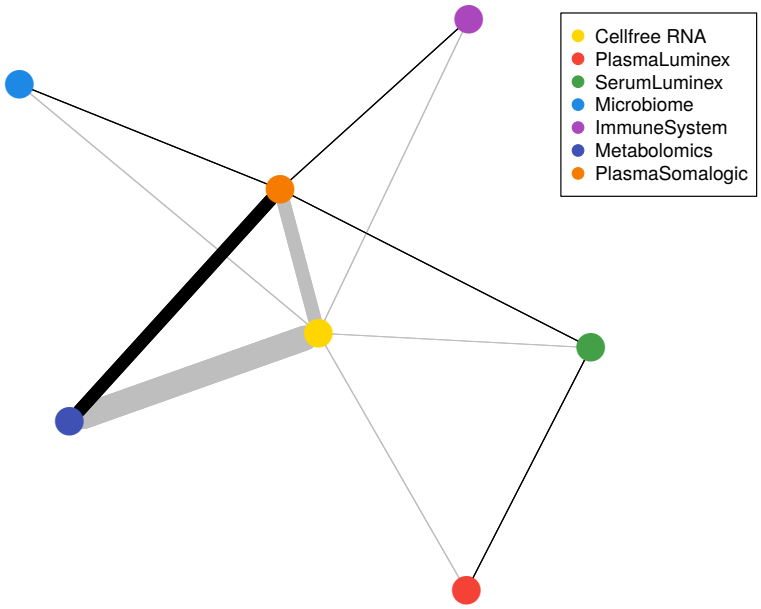
(a) Rank correlation across PlasmaLuminex and PlasmaSomalogic



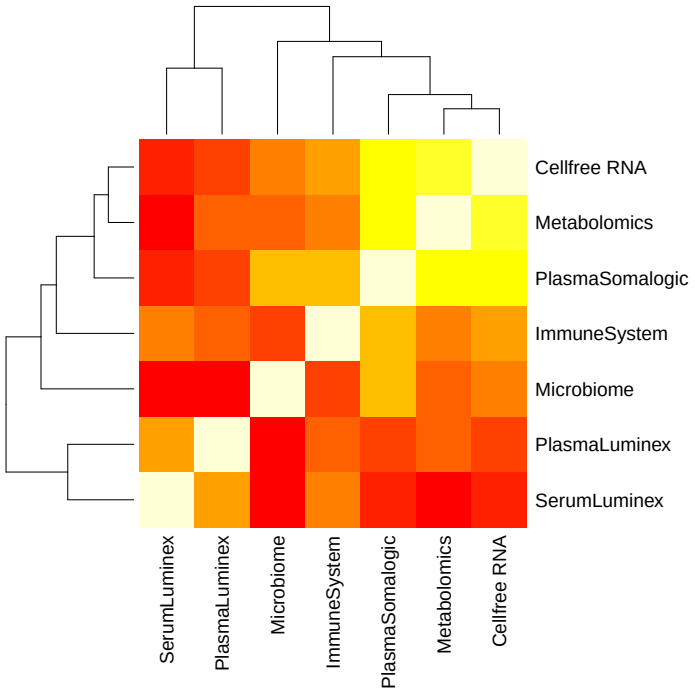
(b) Rank correlation across SerumLuminex and PlasmaSomalogic

Figure 6.12 Illustration of rank correlation among a number of datasets cont. Left panel shows the network representation of RGCCA after Bonferroni adjustment such that presence of an edge between a pair of nodes shows strong correlation between those nodes. Right panel simply demonstrates the heatmap visualization of correlation among two datasets.

Regularized generalized canonical correlation analysis (RGCCA) is the second unsupervised method to study the relationships across the datasets (Tenenhaus and Tenenhaus, 2014). This method combines the advantages of multi-block data analysis and the flexibility of partial least squares regression for unraveling the ties among the different variables of the given multiomics dataset. The result of this algorithm is shown in Figure 6.13(a) as network visualization of multiomics dataset. Thickness of an edge shows the strength and the color shows the direction of correlation black+, and gray−. Figure 6.13(b) displays the heatmap visualization of multivariate correlation among all multiomics dataset. Since Serum and Plasma generated from Luminex family are the most similar omics, they are grouped in one cluster. The remaining datasets show more consistency by forming another tangible cluster.



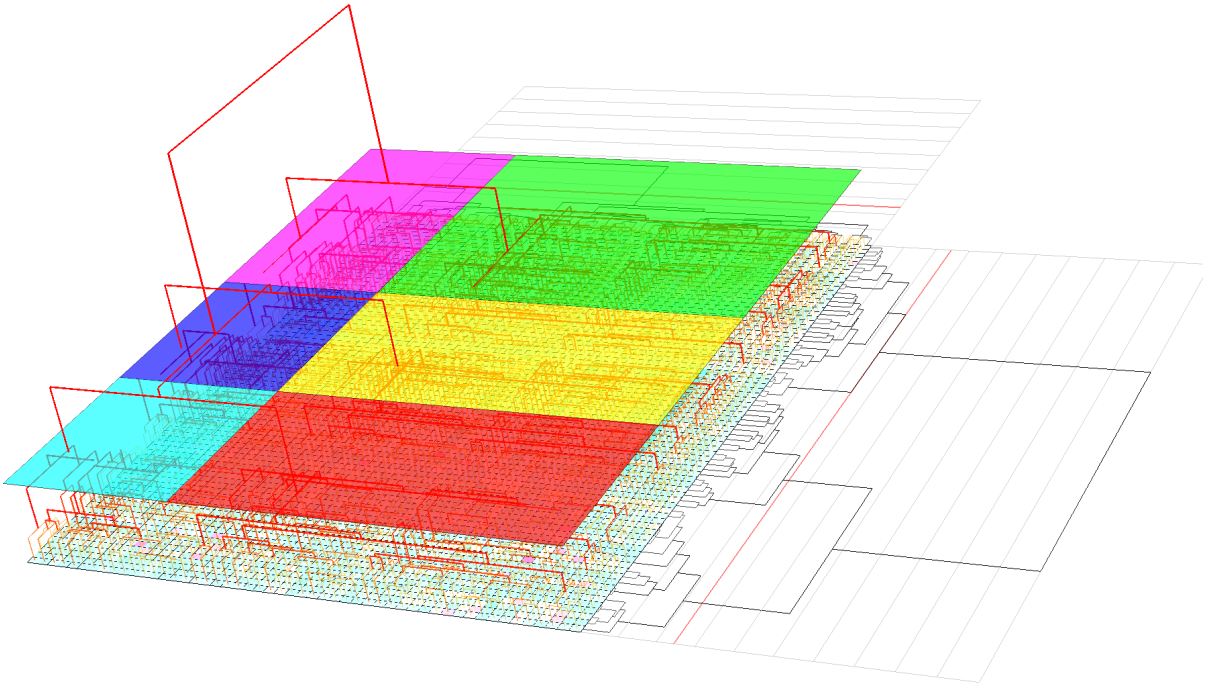
(a) Network visualization of multiomics dataset. Thickness of an edge shows the strength and the color shows the direction of correlation black+, and gray−.



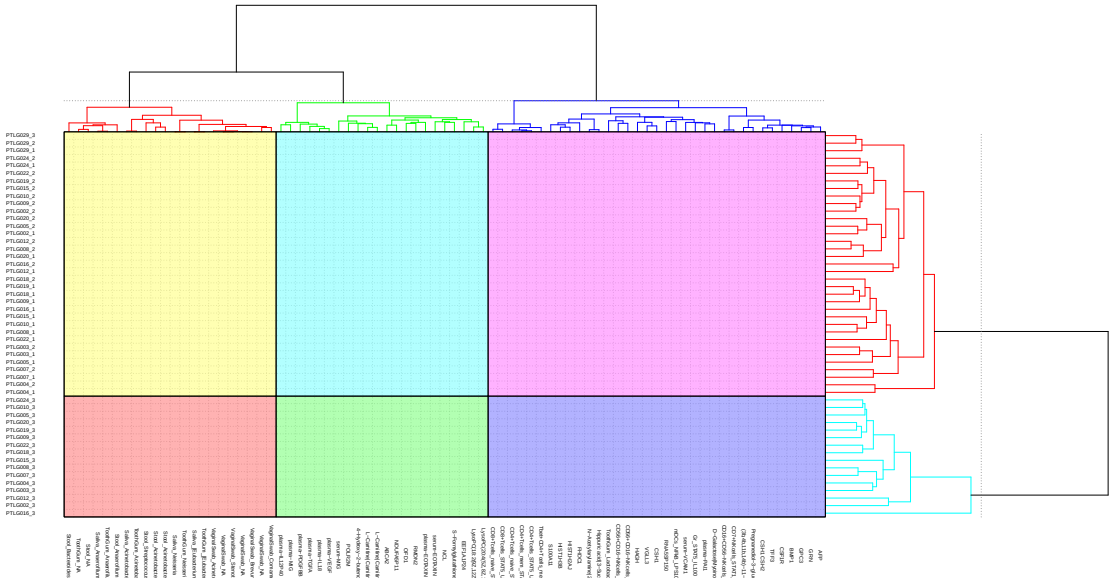
(b) Heatmap visualization of multivariate correlation among all multiomics dataset

Figure 6.13 Unsupervised RGCCA performance on the seven multiomics dataset. Since Serum and Plasma generated from Luminex family are the most similar omics, they are grouped in one cluster. The remaining datasets show more consistency by forming another tangible cluster.

None of the two unsupervised approaches for clustering analysis of multiomics dataset are effectively designed to show meaningful pattern underlying the data. In order to extract more information from the data and describe the correlation among multiomics features and patients intuitively, hierarchical biclustering integrative model is performed on this dataset by the proposed forestogram. Figure 6.14 shows the result of this integrative modeling in two different visualization graphs. Forestogram representation in Figure 6.14(a), depicts the 3D model of hierarchical combination of features and associative patients with 6 automatically selected biclusters. In addition to the 3D model, the 2D projection also demonstrates the biclusters found by the FORIC model selection technique with features and patients label shown in Figure 6.14(b). Furthermore, the same row and column dendrograms extracted from the 2D projection are taken to show the heatmap in Figure 6.15 for fine-grained correlation of the features combined from the seven datasets. Despite this model integrates the features without the supervision information for patients, it turns out the third trimester of pregnancy is distinguished as a separate bicluster from the first two trimester that are located in one bicluster. Moreover, if three clusters are chosen on row, the first two trimesters are still identifiable up to an acceptable noise level. More interestingly, microbiome features are condensed in a unit of bicluster except one measurement that is merged with immune system features. Similarly, one may separate the subclusters of features generated from the same datasets in the lower level of the forestogram. The larger biclusters show a set of merged features which provides more biological insights for further clinical studies.



(a) Hierarchical biclustering integrative model with forestogram in 3D



(b) Hierarchical biclustering integrative model with forestogram in 2D

Figure 6.14 Unsupervised integrative model through forestogram biclustering on the features of multiomics dataset.

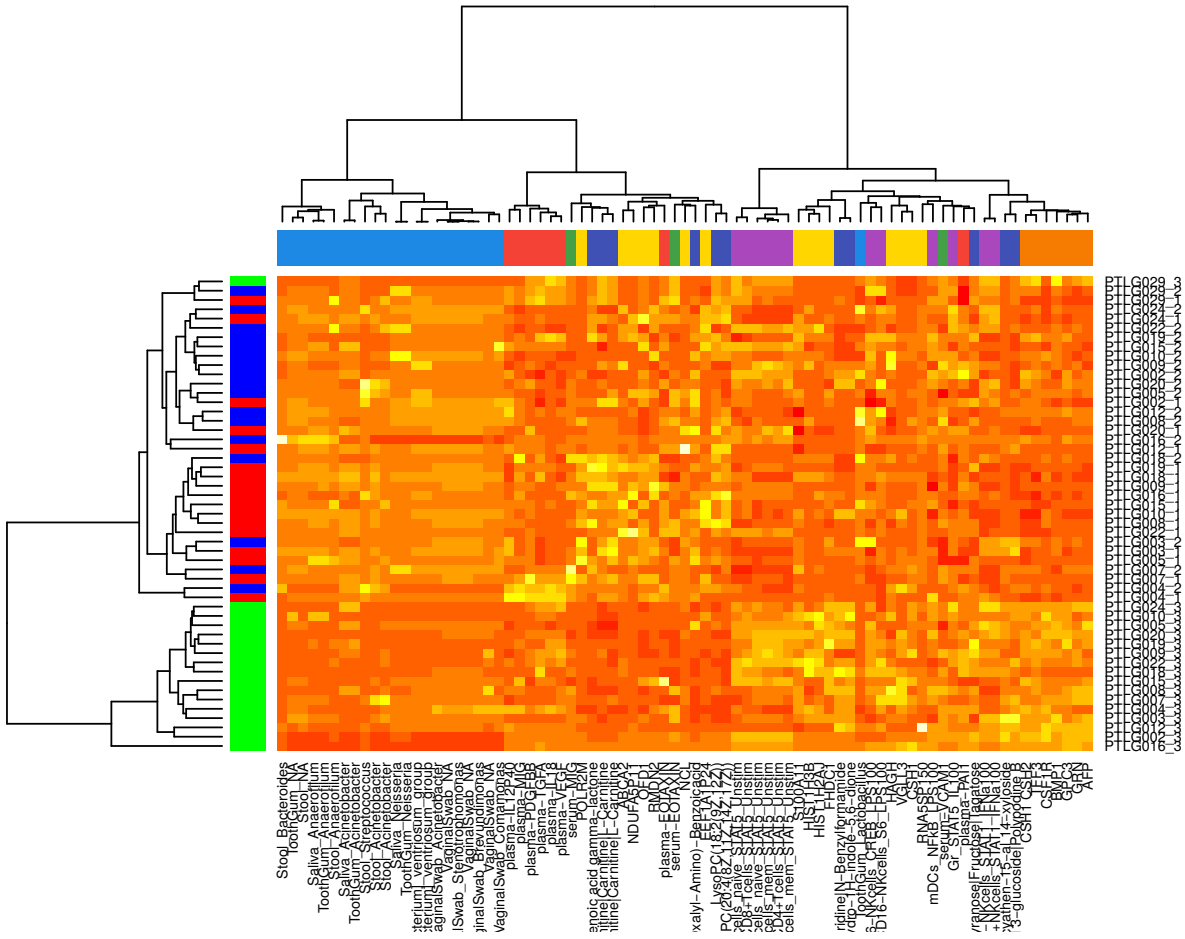


Figure 6.15 Hierarchical biclustering integrative model shown on heatmap.

6.6 Discussion

We described an analysis of seven different biological modalities during term pregnancy. A machine learning approach is used to evaluate the predictive power of each dataset. An additional step is used to combine these predictions to further increase the predictive power. Importantly, these datasets differed in both size and modularity. By taking this two step approach, we prevented larger datasets from overwhelming the final model. This increases both the predictive power, and also the biological interpretation.

Using this approach, we analyze the estimation of the gestational age of the fetus at the time of each sampling. The stacked generalization algorithm produced models more accurate than any individual dataset. Ablation analysis is used to determine the number of omics dataset required for each model, and the impact of each dataset on the final predictions. Importantly, this analysis showed that by retraining the stacked generalization model, other

datasets could partially compensate for the removal of a given dataset. This lays the foundation for analysis of sampling, and for assay costs to strike a cost/predictive-power trade-off in resource-poor settings. Using piece-wise regression and sequential feature reduction, each model is reduced to a limited number of required measurements. These reduced features, then, are used for correlation analysis and visualization.

The approach provided an integrated model of maternal adaptations to pregnancy, highlighting the interconnectivity of multiple biological systems. Notably, strong correlations between metabolomic, proteomic, transcriptomic features and specific immune cell signaling responses pointed at biologically plausible interactions. For example, the model identified a strong relationship between the steroid hormone pregnanolone and the signaling behavior of mDCs and Tregs. mDCs and Tregs play a critical role in feto-maternal tolerance and the maintenance of pregnancy (Erlebacher, 2013; Aluvihare *et al.*, 2004). Our data provide the basis for a novel hypothesis that pregnanolone plays a role in regulation of the function of these two cell types during pregnancy. Alternatively, recent evidence indicating that T-cell can produce pregnenolone, the precursor of pregnanolone (Mahata *et al.*, 2014), suggests immune cells may be a cellular source of pregnanolone production, providing another hypothesis for the observed correlations.

The study also shows that the biological interpretation of observed interactions between two model components benefits from exploring the community of features that strongly correlate with these model components. As such, the integrative model revealed a strong interaction between the proteomic factor CSH1 and STAT5 activity in CD4+ T-cell. However, a community of proteomic factors correlating with CSH1 contained the cytokine IL-2, a canonical activator of the JAK/STAT5 signaling pathway in CD4+ T-cell (Mahmud *et al.*, 2013). Together with our in vitro data showing that stimulation with IL-2 but not with CSH1 results in STAT5 phosphorylation in CD4+ T-cell, these findings suggest that the interaction between CSH1 and STAT5 activity in CD4+ T-cell is likely indirectly mediated by IL-2. For example, activation of the PRL/CSH1 receptor in cells other than T-lymphocytes has been shown to promote the transcription of IL-2 (Sun *et al.*, 2004). CSH1 may thus be implicated in the paracrine regulation of T-cell function through positive regulation of IL-2 gene expression in other immune or non-immune cell types.

A two-layer cross-validation procedure is used in this analysis. The inner layer enables optimization for the hyper parameters of the elastic net model. The outer layer ensures the generalizability of the results to the previously unseen samples. To increase sample size, each sample extracted at a trimester from a single subject is treated as an independent data point. To ensure the models are not biased by the dependency between samples donated by the same subject, all three trimesters of a given subject are excluded together in the same

cross-validation fold. Therefore, the reported results are based on models that have access to no samples from a subject in the test-set. The samples used for testing purposes in all cross-validation layers are synchronized across all models. Therefore, all test-set results (including those of the stacked generalization models) are reported only on samples that are blinded to all previous analysis.

This study has several limitations that have inspired our future plans. First, the cohort size for this proof-of-concept study is relatively small and recruitment from a single-care center limited the diversity of the dataset. Despite this, we are able to capture the chronology of biological changes during pregnancy. However, given the racial disparities in pregnancy outcomes, replicating this analysis in more diverse cohorts is crucial. We have engaged in several multi-national collaborations to directly address this. Second, the number of measurements is significantly larger than the cohort size, which increased the possibility of false positives. In addition to carefully designed cross-validation, feature reduction and clustering Bien and Tibshirani (2011); Witten and Tibshirani (2010); Partovi Nia and Davison (2015) can be used to improve the predictive power of multivariate models in high-dimensional settings. Finally, the current dataset included only one sample per trimester. In the future, high-resolution sampling together with linear mixed effect models Gałecki and Burzykowski (2013) can produce increasingly more accurate estimation of pregnancy related events (including onset of labor) using serial sampling throughout pregnancy.

In summary, our study revealed a precisely timed chronology of responses over the course of term pregnancy. This is enabled using seven high-throughput longitudinal biological assays of the same patient cohort. The computational pipeline produced can increase predictive power by combining datasets of various sizes and modularities in a balanced way. We expect this pipeline to be applicable to a wide range of studies beyond the field of pregnancy. Similarly, the dataset produced here provides a unique resource for future biological investigations. Particularly, this pipeline can be used to identify correlates of any other features from one of the seven datasets that may be identified in future studies. Finally, by characterizing the biological chronology of normal pregnancy, this study provides the conceptual backbone and analytical framework to analyze the complex interplays between various biological modalities that govern preterm birth and other pregnancy-related pathologies.

CHAPTER 7 GENERAL DISCUSSION

We present a new statistical methodology for biclustering problem relying on the hierarchical manner of agglomerative approach for combining pairs of biclusters iteratively towards building the suggested forestogram. This statistical framework demonstrates the hierarchical evolution of biclusters with an intuitive visualization scheme. Additionally, we show how the number of biclusters can be revealed automatically with FORIC as a subtle connection from Bayesian model-based view to the hierarchical nature of forestogram. This statistical setting for the suggested Bayesian model provides enough flexibility supported by data to incorporate a powerful model for applying statistical inference on the data.

For the purpose of performance analysis and applicability, we directly perform this new framework for biclustering on two different applications: 1) in public transport, and 2) in bioinformatics. In the following we briefly discuss the forestogram framework's properties and practical usages for grouping blocks of rows and columns that are interrelated with each other. Then we argue why this approach is promising for the applications that are introduced in this thesis. We complete this discussion by mentioning the limitations of the suggested method and the mitigations to relieve the drawbacks.

7.1 Forestogram

Despite very simple assumptions that forestogram framework is built on, the powerful machinery behind this model-based biclustering approach, makes the framework rich enough to detect broad range of patterns that do not necessarily come from the model. There are a number of properties that make this happens for real data. First of all, hierarchical approach uses Euclidean measure to define the dissimilarity analogous to many data analysis methods. The second interesting property is the connection of this model to the well-known clustering technique i.e k-means (Hartigan and Wong, 1979). In this regard, the hierarchical clustering with Ward's linkage is shown to be equivalent to k-means cost in order to produce the groups of data with spherical shape (Telgarsky and Dasgupta, 2012). However, in order to incorporate other shape of biclusters, the squared Euclidean distance can be changed into Bregman divergence–a general class of distortion functions with connection to the exponential family (Banerjee *et al.*, 2005, 2007). Separability condition ensures, under mild condition meaningful biclusters can be shown on the forestogram. Furthermore, FORIC is suggested to quantify the number of biclusters on the forestogram. The fast computation of clusters' dissimilarity at each iteration is feasible by Lance-William property (Lance and Williams,

1966) to speed up the algorithm. Consequently, putting all these components together, creates a unified framework to perform and analyze the biclustering task through forestogram.

In order to show the efficiency of this framework, we investigate the performance of forestogram on two applied domains.

7.2 Public transport

The main purpose of public transport data analysis in this thesis is to segmenting the similar users into subclusters based on two distinct information gathered from smart card data, known as spatial-temporal data. First of all, we consider the temporal data separately by introducing the semi-circle projection to analyze this information so as to unraveling the hourly patterns in the data. Furthermore, we treat the temporal data as a latent variable to use the Euclidean measure as a geodesic metric on spatial used locations to employ forestogram for extracting the similar spatial-temporal pattern across the users.

In this study we discretized the continuous time usage into hourly binary vector to define the similarity among the users' behavior. We also take Cartesian coordinates for encoding the geographical locations. Using temporal sensitivity analysis by preprocessing the hourly usages instead of equidistant intervals, a dynamic discretization can conduct a better measure of representation for the temporal behavior in the network. Devising new methods for continuous time data based on time series algorithm, using polar coordinate for explaining the location information according to downtown centrality are the next steps that we would like to inspect for modeling the public transport data. In addition, card-day usage or monthly usage could be carried out separately according to the purpose of data analysis to see how clusters vary based on different representation of the temporal information. Furthermore, the temporal data analysis introduced in this study is flexible enough so that by changing the dissimilarity measure a similar analysis can be investigated for temporal behavior in public transit network.

In order to incorporate the temporal information with collection of spatial coordinates, alternative approaches could be studied as well. For instance, time of the entrance can be embedded as a weight to the geographical location. Another scenario is to treat both information independently in the first place, such that by tuning the contribution of each component through convex combination of two parts, the unified model can be amounted to.

For the future direction of this research, we are working on new methods to visualize the spatial-temporal patterns to be comprehensible more intuitively for the transport practitioners. For the spatial analysis of clustering we can use other measure of overlapping between two routes that are relevant for the spatial component of the clustering or alternatives for model estimation which requires to be independent. For the distribution of travel demand,

we may observe other distribution for the usage of bus stops. Moreover, the new overwhelming trend of car pooling technology has raised new challenges for transportation so that integrating public transport data with car pooling, bike sharing, and other similar systems can help better serving society by minimizing the cost and maximizing the quality of the traffic.

7.3 Bioinformatics

Gestational age prediction during term pregnancy is vital toward preventing preterm birth as a second cause for child's death. Moreover, by increasing the accuracy of delivery prediction, hospitals can better schedule their clinical system in order to maximizing the throughput of the health care system and improving the service quality. Mechanism of pregnancy is achieved by interaction between different biological modalities so that we gather a multiomic dataset including measurements from the immunome, transcriptome, microbiome, proteome, and metabolome to predict the gestational age with. In addition to each dataset alone, we suggest two integrative models for multiomics data integration analysis with and without gestational age target. In the supervised integrative fashion, stacked generalization model is used based on cross-validated elastic net that can increase the predictive power by combining datasets of various sizes and modularities in a balanced way.

The unsupervised version provides an extensive model to a context free integrative model to be used for an arbitrary clinical analysis regardless of the supervised labels. However, to some extent, it shows that meaningful features are taken by this model even for pregnancy case. Furthermore, the forestogram visualization provides an intuitive interpretation of features to demonstrate how they combine together to build the integrative model.

This study has several limitations that have inspired our future plans. Despite, the cohort size for this study that is relatively small and the limited diversity of the patients in dataset, we were able to capture the chronology of biological changes during pregnancy. However, given the racial disparities in pregnancy outcomes, replicating this analysis in more diverse cohorts is crucial. We deal with several multi-national collaborations to directly address this. Second, prediction accuracy dropped for large omic datasets due to the underdetermined matrix of the data where number of measurements is significantly larger than the cohort size. In addition to carefully designed cross-validation, feature reduction and semi-supervised biclustering can be used to improve the predictive power of multivariate models in high-dimensional settings. Finally, the current dataset included only one sample per trimester.

In the future, high-resolution sampling together with linear mixed effect models suggested by Gałecki and Burzykowski (2013), can produce increasingly more accurate estimation of

pregnancy related events using serial sampling throughout pregnancy.

7.4 Limitations of forestogram and hierarchical algorithms

Clustering is an NP-hard problem. However, in order to compute a feasible guess an approximation is made to the exact optimal solution up to certain level of error. Hierarchical approaches are suboptimal solutions to clustering that are highly prone to converge to local optimal solution due to the greedy nature of the algorithm. Ackermann *et al.* (2014) show that agglomerative complete linkage clustering is an $\mathcal{O}(\log k)$-approximation to the k-clustering problem. For general hierarchical clustering, for every k, the approximation factor of k-clustering is at most eight times of the optimal solution (Dasgupta and Long, 2005). Regardless of the intrinsic limitation of hierarchical approaches, they provide a nice visualization with a partitioning of data for different number of clusters that vary from 1 to n. Moreover, changing the linkage function makes this approach flexible to produce different type of clusters with interesting properties, though naïve computation of these methods could be inefficient for moderately large scale datasets.

In our suggested hierarchical framework, Lance-William speed up technique accelerates the dissimilarity computation at each step, but the overall time complexity is still cubic. This can be reduced to $n^2 \log n$ using the priority queue (Schrage, 1967) or even square with reciprocal nearest neighbor introduced in Murtagh and Contreras (2011). On the other hand, the normal distribution assumption with the same variance for all biclusters, limits the FORIC for number of biclusters prediction with the cost of fast computation of predictive distribution. Additionally, due to the cubic time complexity and square space complexity performing forestogram on big data on personal computers is cumbersome. Even if the algorithm is performed on big data, it is not possible to illustrate a large forestogram object to elucidate all rows and columns details and labels.

Moreover, in certain problems the data is given in terms of tensor structure including replicates of the same matrix data type with different measurements. The suggested forestogram is not tailored to address the biclustering of the 3D matrix data structure. In this regard, if one wants to use this framework, the tensor should be converted to a bigger matrix or with some statistical methods a matrix representative for the given tensor requires to use forestogram for biclustering.

7.5 Improvement

Our proposed forestogram model can be extended in various directions while the focus of this research was to develop an efficient framework for fast computing hierarchical biclustering

model. The configuration of current setting is constructed by distanced-based hierarchical approach for the sake of tractable computation of forestogram. However, it is possible to merge the biclusters according to a probabilistic model of the data with statistical measure e.g. Dirichlet mixture (Heller and Ghahramani, 2005), and maximum likelihood generative tree (Castro *et al.*, 2004). Other linkages such as minmax introduced in Bien and Tibshirani (2011), and robust linkage (Balcan *et al.*, 2014) could also be implemented for forestogram if heavy time complexity of dissimilarity update does not matter at each iteration. Moreover, for predicting the number of biclusters we use a uniform prior on likelihood to implement the FORIC effectively, while other priors such as multinomial-Dirichlet distribution can also be plugged in the predictive distribution in favor of small clusters (Heard *et al.*, 2006). The next issue of biclustering is big data trend, to this end, one can run k-means with plenty of clusters in order to invoke forestogram with Ward's linkage on top of the extracted cluster centers since Ward's linkage is shown to be equivalent to k-means clustering in Telgarsky and Dasgupta (2012). Since, k-means is sensitive to initial cluster centers, we suggest to utilize the stochastic variant of k-means with no means parameter developed in Partovi Nia *et al.* (2017). The stochastic scheme is promising in terms of parallel implementation especially for GPU parallel computing in order to scale up the computation for big chunk of data. We already know that under Ward's linkage our forestogram is equivalent to apply k-means at each iteration, because they optimize the same cost function. In addition, we are interested to investigate the potential theoretical connections of this methodology with other clustering and biclustering approaches such as Spectral clustering, nonnegative matrix factorization, Gaussian mixture, convex clustering, etc. in the future.

CHAPTER 8 CONCLUSION

The importance of unsupervised learning, such as clustering and biclustering is unprecedented in the advent of modern data analysis era. There is a number of state-of-the-art methods for clustering that are categorized into two major branches, hierarchical and partitional. Partitional approach is effective when the number of clusters is already known. For many real world data, identifying this information is part of the data analysis study to realize more knowledge from the data. Toward this goal, model-based hierarchical approach is proposed in this thesis such that flexibility of the model is led by the support of data to incorporate a powerful model for carrying out statistical inference on the data. The other advantage of this perspective is the visualization property to demonstrate the model as an intuitive way of describing the hidden structures underlying the data. This conceptual view in conjunction with revealing the biclusters automatically make an autonomous framework for biclustering analysis of an arbitrary data matrix.

The main contribution of this research is designing and developing the forestogram methodology to biclustering analysis of a data matrix without supervised information in general. In this regard, we propose a hierarchical approach to address this problem in particular. Moreover, we implement FORIC by which the number of biclusters is determined automatically. Consequently, the evolution of biclusters is demonstrated by the extension of dendrogram, that we call forestogram as a visual representation for recursive merged biclusters through row or column. In order to validate the efficiency of forestogram in applied fields, we suggest to perform this framework on public transport, and bioinformatics.

User's behavior modeling plays a central role in studying and analyzing public transport data collected from smart card. In order to uncover the spatial-temporal behavior of subscribers in the public transit network, we suggest a smart projection to map the binary vector of timestamped data into a 3D space. This SCP conserves the pairwise similarity among the users the same in the new space with a visualization guide for better understanding of the temporal pattern. Seventeen clusters are identified in terms of single trip, regular users, late commuters, long day, midday, active and inactive groups as the temporal behavior of the users by applying agglomerative hierarchical clustering on the transformed data.

Furthermore, under the light of forestogram development, we suggest a new perspective to extract the spatial patterns through temporal latent variable. Spatial data contains worthwhile information about the geographical details of each bus stop and are stored sequentially following the order of temporal usage. Finding an appropriate measure of similarity for spatial-temporal behaviors in public transit network is challenging due to different scenarios

that may arise in to express the similarity among a pair of spatial trajectories. In this regard, a customized version of forestogram is designed to extract the similar group of users with their corresponding temporal and spatial patterns simultaneously. To this end, the spatial coordinate of GPS location (x, y) from the corresponding time interval takes the same place in the data matrix. This way, we can perform the Euclidean distance to the entries of the matrix while the latent time information is implicitly taken into account

In the analysis of pregnancy, our empirical study of multiomics features reveals a chronology of biologically diverse events over the course of pregnancy. This is enabled using seven high-throughput longitudinal biological assays of the same patient cohort. The computational pipeline introduced in this thesis can increase predictive power by combining datasets of various sizes and modularities in a balanced way. Moreover, by performing the suggested forestogram for biclustering on biological chronology of measured features, we can provide an unsupervised model for pregnancy as well. Our biclustering framework presents an analytical unsupervised model for the complex interplays between various biological modalities that govern preterm birth and other pregnancy-related pathologies.

REFERENCES

Margareta Ackerman and Shai Ben-david (2009). Clusterability: A theoretical study. D. Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS-09)*. Journal of Machine Learning Research - Proceedings Track, vol. 5, 1–8.

Ackermann, Marcel R and Blömer, Johannes and Kuntze, Daniel and Sohler, Christian (2014). Analysis of agglomerative clustering. *Algorithmica*, *69*(1), 184–215.

Agard, Bruno and Morency, Catherine and Trépanier, Martin (2008). Mining smart card data from an urban transit network. J. Wang, editor, *Encyclopedia of Data Warehousing and Mining*, IGI Global. 1292–1302.

Saeed Aghabozorgi and Ali Seyed Shirkhorshidi and Teh Ying Wah (2015). Time-series clustering – a decade review. *Information Systems*, *53*, 16–38.

Aghaeepour, Nima and Ganio, Edward A. and Mcilwain, David and Tsai, Amy S. and Tingle, Martha and Van Gassen, Sofie and Gaudilliere, Dyani K. and Baca, Quentin and McNeil, Leslie and Okada, Robin and Ghaemi, Mohammad S. and Furman, David and Wong, Ronald J. and Winn, Virginia D. and Druzin, Maurice L. and El-Sayed, Yaser Y. and Quaintance, Cecele and Gibbs, Ronald and Darmstadt, Gary L. and Shaw, Gary M. and Stevenson, David K. and Tibshirani, Robert and Nolan, Garry P. and Lewis, David B. and Angst, Martin S. and Gaudilliere, Brice (2017). An immune clock of human pregnancy. *Science Immunology*, *2*(15), eaan2946.

Aghaeepour, Nima and Hoos, Holger H (2013). Ensemble-based prediction of rna secondary structures. *BMC Bioinformatics*, *14*(1), 139.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. B. N. Petrov and F. Csaki, editors, *Second International Symposium on InformationTheory*. 267–281.

Akavia, Uri David and Litvin, Oren and Kim, Jessica and Sanchez-Garcia, Felix and Kotliar, Dylan and Causton, Helen C and Pochanard, Panisa and Mozes, Eyal and Garraway, Levi A and Peter, Dana (2010). An integrated approach to uncover drivers of cancer. *Cell*, *143*(6), 1005–1017.

Ali, Atizaz and Kim, Jooyoung and Lee, Seungjae (2016). Travel behavior analysis using smart card data. *KSCE Journal of Civil Engineering*, *20*(4), 1532–1539.

Alqadah, Faris and Reddy, Chandan K and Hu, Junling and Alqadah, Hatim F (2015). Biclustering neighborhood-based collaborative filtering method for top-n recommender systems. *Knowledge and Information Systems*, *44*(2), 475–491.

Azalden Alsger and Behrang Assemi and Mahmoud Mesbah and Luis Ferreira (2016). Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C Emerging Technologies*, *68*, 490–506.

Aluvihare, Varuna R and Kallikourdis, Marinos and Betz, Alexander G (2004). Regulatory t cells mediate maternal tolerance to the fetus. *Nature Immunology*, *5*(3), 266–271.

Berk Anbaroglu and Benjamin Heydecker and Tao Cheng (2014). Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies*, *48*, 47–65.

Arora, Sanjeev and Ge, Rong and Kannan, Ravindran and Moitra, Ankur (2012). Computing a nonnegative matrix factorization – provably. *Proceedings of the 44th symposium on Theory of Computing*. ACM, STOC '12, 145–162.

Balcan, Maria-Florina and Liang, Yingyu and Gupta, Pramod (2014). Robust hierarchical clustering. *The Journal of Machine Learning Research*, *15*(1), 3831–3871.

Banerjee, Arindam and Dhillon, Inderjit and Ghosh, Joydeep and Merugu, Srujana and Modha, Dharmendra S (2007). A generalized maximum entropy approach to bregman coclustering and matrix approximation. *Journal of Machine Learning Research*, *8*, 1919–1986.

Banerjee, Arindam and Merugu, Srujana and Dhillon, Inderjit S and Ghosh, Joydeep (2005). Clustering with bregman divergences. *Journal of Machine Learning Research*, *6*, 1705–1749.

Ben-Dor, Amir and Shamir, Ron and Yakhini, Zohar (1999). Clustering gene expression patterns. *Journal of Computational Biology*, *6*(3-4), 281–297.

Bendall, Sean C and Simonds, Erin F and Qiu, Peng and El-ad, D Amir and Krutzik, Peter O and Finck, Rachel and Bruggner, Robert V and Melamed, Rachel and Trejo, Angelica and Ornatsky, Olga I and others (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, *332*(6030), 687–696.

Bien, Jacob and Tibshirani, Robert (2011). Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, *106*(495), 1075–1084.

Bordagaray, Maria and dell'Olio, Luigi and Ibeas, Angel and Cecín, Patricia (2014). Modelling user perception of bus transit quality considering user and service heterogeneity. *Transportmetrica A Transport Science*, *10*(8), 705–721.

Borg, Ingwer and Groenen, Patrick J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics)*. Springer, second edition.

Borgwardt, Karsten M and Ong, Cheng Soon and Schönauer, Stefan and Vishwanathan, SVN and Smola, Alex J and Kriegel, Hans-Peter (2005). Protein function prediction via graph kernels. *Bioinformatics*, *21*(1), 47–56.

Boutsinas, Basilis (2013). Machine-part cell formation using biclustering. *European Journal of Operational Research*, *230*(3), 563–572.

Breiman, Leo (1996). Stacked regressions. *Machine learning*, *24*(1), 49–64.

Busygin, Stanislav and Prokopyev, Oleg and Pardalos, Panos M. (2008). Biclustering in data mining. *Computers & Operations Research*, *35*(9), 2964–2987.

Deng Cai and Xiaofei He and Xiaoyun Wu and Jiawei Han (2008). Non-negative matrix factorization on manifold. *International Conference on Data Mining (ICDM'08)*.

Callahan, Benjamin J and DiGiulio, Daniel B and Goltsman, Daniela S Aliaga and Sun, Christine L and Costello, Elizabeth K and Jeganathan, Pratheepa and Biggio, Joseph R and Wong, Ronald J and Druzin, Maurice L and Shaw, Gary M and others (2017). Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of us women. *Proceedings of the National Academy of Sciences*, *114*(37), 9966–9971.

Carel, L. and Alquier, P. (2017). Simultaneous Dimension Reduction and Clustering via the NMF-EM Algorithm. *ArXiv e-prints arxiv: 1709.03346*.

Carlsson, Gunnar and Mémoli, Facundo (2010). Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, *11*, 1425–1470.

Castro, Rui M and Coates, Mark J and Nowak, Robert D (2004). Likelihood based hierarchical clustering. *IEEE Transactions on Signal Processing*, *52*(8), 2308–2321.

Malika Charrad and Nadia Ghazzali and Véronique Boiteau and Azam Niknafs (2014). NbClust an R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36.

Chen, Gary K and Chi, Eric C and Ranola, John Michael O and Lange, Kenneth (2015). Convex Clustering: An Attractive Alternative to Hierarchical Clustering. *PLoS Computational Biology*, *11*(5), e1004228.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. *International Conference on Intelligent Systems for Molecular Biology*. vol. 8, 93–103.

Tak-chung Fu (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164–181.

G. Claeskens and N. L. Hjort (2008). *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Clarke, Robert and Ressom, Habtom W and Wang, Antai and Xuan, Jianhua and Liu, Minetta C and Gehan, Edmund A and Wang, Yue (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, *8*(1), 37.

Conesa, Ana and Madrigal, Pedro and Tarazona, Sonia and Gomez-Cabrero, David and Cervera, Alejandra and McPherson, Andrew and Szcześniak, Michał Wojciech and Gaffney, Daniel J and Elo, Laura L and Zhang, Xuegong and others (2016). A survey of best practices for rna-seq data analysis. *Genome Biology*, *17*(1), 1–19.

Dasgupta, Sanjoy and Long, Philip M (2005). Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, *70*(4), 555–569.

de Oña, Rocio and de Oña, Juan (2015). Analysis of transit quality of service through segmentation and classification tree techniques. *Transportmetrica A Transport Science*, *11*(5), 365–387.

Del Castillo, JM and Benitez, FG (2013). Determining a public transport satisfaction index from user surveys. *Transportmetrica A Transport Science*, *9*(8), 713–741.

Dethlefsen, Les and McFall-Ngai, Margaret and Relman, David A (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, *449*(7164), 811–818.

Dey, Kushal K and Hsiao, Chiaowen Joyce and Stephens, Matthew (2017). Visualizing the structure of rna-seq expression data using grade of membership models. *PLoS Genetics*, *13*(3), e1006599.

DiGiulio, Daniel B and Callahan, Benjamin J and McMurdie, Paul J and Costello, Elizabeth K and Lyell, Deirdre J and Robaczewska, Anna and Sun, Christine L and Goltsman, Daniela SA and Wong, Ronald J and Shaw, Gary and others (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences*, *112*(35), 11060–11065.

Chris Ding and Xiaofeng He and Horst D. Simon (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. *in SIAM International Conference on Data Mining*. 606–610.

David Donoho and Victoria Stodden (2004). When does non-negative matrix factorization give a correct decomposition into parts ? S. Thrun, L. K. Saul and P. B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, MIT Press. 1141–1148.

Dou, Mengyu and He, Tieke and Yin, Hongzhi and Zhou, Xiaofang and Chen, Zhenyu and Luo, Bin (2015). *Predicting Passengers in Public Transportation Using Smart Card Data*, Springer International Publishing, Cham. 28–40.

Druckmann, René and Druckmann, Marc-Alexandre (2005). Progesterone and the immunology of pregnancy. *The Journal of Steroid Biochemistry and Molecular Biology*, *97*(5), 389–396.

Eisen, Michael B and Spellman, Paul T and Brown, Patrick O and Botstein, David (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868. Open source C library for clustering.

Justin Eldridge and Mikhail Belkin and Yusu Wang (2015). Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. P. Grünwald, E. Hazan and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*. PMLR, Paris, France, vol. 40 of *Proceedings of Machine Learning Research*, 588–606.

Emilsson, Valur and Thorleifsson, Gudmar and Zhang, Bin and Leonardson, Amy S and Zink, Florian and Zhu, Jun and Carlson, Sonia and Helgason, Agnar and Walters, G Bragi and Gunnarsdottir, Steinunn and others (2008). Genetics of gene expression and its effect on disease. *Nature*, *452*(7186), 423–428.

Eren, Kemal and Deveci, Mehmet and Küçüktunç, Onur and Çatalyürek, Ümit V (2013). A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, *14*(3), 279–292.

Erlebacher, Adrian (2013). Immunology of the maternal-fetal interface. *Annual Review of Immunology*, *31*, 387–411.

B. Everitt and S. Landau and M. Leese and D. Stahl (2011). *Cluster Analysis*. Wiley Publishing, New York, fifth edition.

Farrar, David B (2006). *Some Model-Based and Distance-Based Clustering Methods for Characterization of Regional Ecological Stressor-Response Patterns and Regional Environmental Quality Trends*. doctoral thesis, Virginia Polytechnic Institute and State University.

Feng, Chen-Chieh and Wang, Yi-Chen and Chen, Chih-Yuan (2014). Combining geo-som and hierarchical clustering to explore geospatial data. *Transactions in GIS*, *18*(1), 125–146.

Florek, K and Łukaszewicz, J and Perkal, J and Steinhaus, Hugo and Zubrzycki, S (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicae*. Institute of Mathematics Polish Academy of Sciences, vol. 2, 282–285.

Fowler, Anna and Heard, Nicholas A (2012). On two-way bayesian agglomerative clustering of gene expression data. *Statistical Analysis and Data Mining*, *5*(5), 463–476.

C. Fraley and A. E. Raftery (1999). MCLUST software for model-based cluster analysis. *Journal of Classification*, *16*, 297–306.

Freiria, Susana and Ribeiro, Bernardete and Tavares, Alexandre O (2015). Understanding road network dynamics: Link-based topological patterns. *Journal of Transport Geography*, *46*, 55–66.

Fridley, Brooke L and Lund, Steven and Jenkins, Gregory D and Wang, Liewei (2012). A bayesian integrative genomic model for pathway analysis of complex traits. *Genetic Epidemiology*, *36*(4), 352–359.

Froehlich, Jon and Krumm, John (2008). Route prediction from trip observations. technical report, SAE Technical Paper.

T. Fuse and K. Makimura and T. Nakamura (2012). Observation of travel behavior by ic card data and application to transportation planning. *Proceedings of the ACM Knowledge Discovery from Mobility Data for Intelligent Transportation Systems*. ACM, New York, NY, USA, UrbComp '12, 79–86.

Galba, Tomislav and Balkic, Zoran and Martinovic, Goran (2013). Public transportation bigdata clustering. *International Journal of Electrical and Computer Engineering Systems*, *4*(1.), 21–26.

Gałecki, Andrzej and Burzykowski, Tomasz (2013). Linear mixed-effects model. *Linear Mixed-Effects Models Using R*, Springer. 245–273.

Gallotti, Riccardo and Barthelemy, Marc (2015). The multilayer temporal network of public transport in Great Britain. *Scientific Data*, *2*, 140056.

Gan, Xiangchao and Liew, Alan and Yan, Hong (2008). Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinformatics*, *9*, 209.

Ge, Guangtao and Wong, G William (2008). Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, *9*(1), 275.

Ghaemi, Mohammad Sajjad and Agard, Bruno and Partovi Nia, Vahid (2017a). Forestogram: A Visualization Framework for Hierarchical Biclustering. Manuscript.

Ghaemi, Mohammad Sajjad and Agard, Bruno and Partovi Nia, Vahid and Trépanier, Martin (2015). Challenges in spatial-temporal data analysis targeting public transport. *IFAC-PapersOnLine*, *48*(3), 442–447.

Ghaemi, Mohammad Sajjad and Agard, Bruno and Trépanier, Martin and Partovi Nia, Vahid (2017b). A visual segmentation method for temporal smart card data. *Transportmetrica A: Transport Science*, *13*(5), 381–404.

Ghaemi, Mohammad Sajjad and Partovi Nia, Vahid and Agard, Bruno (2017c). *hbiclust: Fast Bayesian Hierarchical Biclustering and Forestogram*. R package beta version available on R-forge.

Ghaemi, Mohammad Sajjad and Partovi Nia, Vahid and Agard, Bruno and Aghaeepour, Nima (2017d). Multiomics analysis of host response to pregnancy. Manuscript.

M. Ghasemzadeh and B. C. M. Fung and R. Chen and A. Awasthi (2014). Anonymizing trajectory data for passenger flow analysis. *Transportation Research Part C Emerging Technologies (TRC)*, *39*, 63–79.

Nicolas Gillis (2011). *N. Gillis, Nonnegative Matrix Factorization Complexity, Algorithms and Applications*. doctoral thesis, Universit catholique de Louvain.

Konstantinos Gkiotsalitis and Antony Stathopoulos (2015). A utility-maximization model for retrieving users' willingness to travel for participating in activities from big-data. *Transportation Research Part C Emerging Technologies*, *58, Part B*, 265 – 277.

Gouilleux, F and Wakao, H and Mundt, Maren and Groner, B (1994). Prolactin induces phosphorylation of tyr694 of stat5 (mgf), a prerequisite for dna binding and induction of transcription. *The EMBO Journal*, *13*(18), 4361.

Govaert, G. and Nadif, M. (2013). *Co-clustering: Models, algorithms and applications*. Computing Engineering Series. ISTE-Wiley.

Gu, Jiajun and Liu, Jun S. (2008). Bayesian biclustering of gene expression data. *BMC Genomics*, *9*, S4+.

Hartigan, J.A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, *67*, 123–129.

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

S. Hasan and C. M. Schneider and S. V. Ukkusuri and M. C. González (2012). Spatiotemporal patterns of urban human mobility. *Statistical Physics*, *151*(1-2), 304–318.

Hassanzadeh, Hamid Reza and Phan, John H and Wang, May D (2016). A multi-modal graph-based semi-supervised pipeline for predicting cancer survival. *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 184–189.

Hastie, Trevor J. and Tibshirani, Robert J. and Friedman, J. H. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer, New York, second edition.

He, Linna and Yang, Zhihao and Zhao, Zhehuan and Lin, Hongfei and Li, Yanpeng (2013). Extracting drug-drug interaction from the biomedical literature using a stacked generalization-based approach. *PloS one*, *8*(6), 1–12.

N. A. Heard and C. C. Holmes and D. A. Stephens (2006). A quantitative study of gene regulation involved in the immune response of *Anopheline* mosquitoes An application of

Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, *101*(473), 18–29.

Heller, Katherine A. and Ghahramani, Zoubin (2005). Bayesian hierarchical clustering. *Proceedings of the 22Nd International Conference on Machine Learning*. ACM, New York, NY, USA, ICML '05, 297–304.

Juan C. Herrera and Daniel B. Work and Ryan Herring and Xuegang (Jeff) Ban and Quinn Jacobson and Alexandre M. Bayen (2010). Evaluation of traffic data obtained via gps-enabled mobile phones the mobile century field experiment. *Transportation Research Part C Emerging Technologies*, *18*(4), 568 – 583.

Hochreiter, Sepp and Bodenhofer, Ulrich and Heusel, Martin and Mayr, Andreas and Mitterecker, Andreas and Kasim, Adetayo and Khamiakova, Tatsiana and Van Sanden, Suzy and Lin, Dan and Talloen, Willem and others (2010). Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, *26*(12), 1520–1527.

Holzinger, Emily R and Dudek, Scott M and Frase, Alex T and Pendergrass, Sarah A and Ritchie, Marylyn D (2013). Athena: the analysis tool for heritable and environmental network associations. *Bioinformatics*, *30*(5), 698–705.

Huang, Anna (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand Computer Science Research Student Conference, Christchurch, New Zealand*. 49–56.

Huang, Jih-Jeng and Tzeng, Gwo-Hshiung and Ong, Chorng-Shyong (2007). Marketing segmentation using support vector clustering. *Expert Systems with Applications*, *32*(2), 313–317.

Huang, Sijia and Chaudhary, Kumardeep and Garmire, Lana X (2017). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, *8*, 84.

Huber, Wolfgang and Carey, Vincent J and Gentleman, Robert and Anders, Simon and Carlson, Marc and Carvalho, Benilton S and Bravo, Hector Corrada and Davis, Sean and Gatto, Laurent and Girke, Thomas and others (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, *12*(2), 115–121.

Hubert, Lawrence and Arabie, Phipps (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Ideker, Trey and Thorsson, Vesteinn and Ranish, Jeffrey A and Christmas, Rowan and Buhler, Jeremy and Eng, Jimmy K and Bumgarner, Roger and Goodlett, David R and Aebersold, Ruedi and Hood, Leroy (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, *292*(5518), 929–934.

Izenman, Alan Julian (2008). *Modern Multivariate Statistical Techniques*, vol. 1. Springer Publishing Company, Incorporated.

Jain, Anil K. (2010). Data clustering 50 years beyond k-means. *Pattern Recognition Letters*, *31*, 651–666.

Olle Järv and Rein Ahas and Frank Witlox (2014). Understanding monthly variability in human activity spaces a twelve-month study using mobile phone call detail records. *Transportation Research Part C Emerging Technologies*, *38*, 122 – 135.

Ji, Hongkai and Liu, X Shirley (2010). Analyzing'omics data using hierarchical models. *Nature Biotechnology*, *28*(4), 337–340.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254. The first formal hierarchical clustering algorithm.

Sebastian Kaiser and Rodrigo Santamaria and Tatsiana and Khamiakova and Martin Sill and Roberto Theron and Luis Quintales and Friedrich Leisch. (2013). *Biclust BiCluster Algorithms*. R package version 1.0.2.

Kang, Suk-Jo and Liang, Hong-Erh and Reizis, Boris and Locksley, Richard M (2008). Regulation of hierarchical clustering and activation of innate immune cells by dendritic cells. *Immunity*, *29*(5), 819–833.

A. E. Kass and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Katz, Daniel H and Benson, Mark D and Yang, Qiong and Keyes, Michelle J and Shen, Dongxiao and Sinha, Sumita and Morningstar, Jordan E and Ngo, Debby and O'Sullivan, John F and Shi, Xu and others (2016). Unsupervised hierarchical cluster analysis of combined metabolomic and proteomic profiling data sets from participants in a community-based cohort study.

Le Minh Kieu and Ashish Bhaskar and Edward Chung (2014). Transit passenger segmentation using travel regularity mined from smart card transactions data. *Transportation Research Board 93rd Annual Meeting*. Washington, D.C, 123–134.

Kim, Dokyoon and Shin, Hyunjung and Song, Young Soo and Kim, Ju Han (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *Journal of Biomedical Informatics*, *45*(6), 1191–1198.

Jon M. Kleinberg (2003). An impossibility theorem for clustering. S. Becker, S. Thrun and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. MIT Press, 463–470.

Klingenberg, Bradley and Curry, James and Dougherty, Anne (2009). Non-negative matrix factorization ill-posedness and a geometric algorithm. *Pattern Recognition*, (5), 918–928.

Kurauchi, Fumitaka and Schmöcker, Jan-Dirk (2017). *Public Transport Planning with Smart Card Data*. CRC Press.

Kusakabe, Takahiko and Asakura, Yasuo (2014). Behavioural data mining of transit smart card data a data fusion approach. *Transportation Research Part C Emerging Technologies*, *46*, 179–191.

Lackritz, Eve M and Wilson, Christopher B and Guttmacher, Alan E and Howse, Jennifer L and Engmann, Cyril M and Rubens, Craig E and Mason, Elizabeth M and Muglia, Louis J and Gravett, Michael G and Goldenberg, Robert L and others (2013). A solution pathway for preterm birth: accelerating a priority research agenda. *The Lancet Global Health*, *1*(6), 328–330.

G. N. Lance and W. T. Williams (1966). A general theory of classifactory sorting strategies, i. hierarchical systems. *Computer Journal*, *9*, 373–380.

Larranaga, Pedro and Calvo, Borja and Santana, Roberto and Bielza, Concha and Galdiano, Josu and Inza, Iñaki and Lozano, José A and Armañanzas, Rubén and Santafé, Guzmán and Pérez, Aritz and others (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112.

Neal Lathia and Saniul Ahmed and Licia Capra (2012). Measuring the impact of opening the london shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies*, *22*, 88 – 102.

Lathia, Neal and Capra, Licia (2011). How Smart is Your Smartcard Measuring Travel Behaviours, Perceptions, and Incentives. *Proceedings of the 13th International Conference on Ubiquitous Computing*. ACM, New York, NY, USA, 291–300.

Lathia, Neal and Froehlich, Jon and Capra, Licia (2010). Mining public transport usage for personalised intelligent transport systems. *Proceedings of the 2010 IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, ICDM '10, 887–892.

Lazzeroni, Laura and Owen, Art (2002). Plaid models for gene expression data. *Statistica Sinica*, *12*, 61–86.

Lee, Daniel D. and Seung, H. Sebastian (2000). Algorithms for non-negative matrix factorization. *Proceedings of the 13th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS'00, 535–541.

Li, Quannan and Zheng, Yu and Xie, Xing and Chen, Yukun and Liu, Wenyu and Ma, Wei-Ying (2008). Mining user similarity based on location history. *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, New York, NY, USA, GIS '08, 341–3410.

Liiv, Innar (2010). Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *3*(2), 70–91.

Liu, Li and Johnson, Hope L and Cousens, Simon and Perin, Jamie and Scott, Susana and Lawn, Joy E and Rudan, Igor and Campbell, Harry and Cibulskis, Richard and Li, Mengying and others (2012). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *The Lancet*, *379*(9832), 2151–2161.

Ma, Tai-Yu and Gerber, Philippe and Carpentier, Samuel and Klein, Sylvain (2015). Mode choice with latent preference heterogeneity a case study for employees of the eu institutions in luxembourg. *Transportmetrica A Transport Science*, *11*(5), 441–463.

Xiaolei Ma and Yao-Jan Wu and Yinhai Wang and Feng Chen and Jianfeng Liu (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C Emerging Technologies*, *36*, 1–12.

S. C. Madeira and A. L. Oliveira (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on Computational Biology and Bioinformatics*, *1*, 24–45.

Maetschke, Stefan R and Madhamshettiwar, Piyush B and Davis, Melissa J and Ragan, Mark A (2013). Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics*, *15*(2), 195–211.

Mahata, Bidesh and Zhang, Xiuwei and Kolodziejczyk, Aleksandra A and Proserpio, Valentina and Haim-Vilmovsky, Liora and Taylor, Angela E and Hebenstreit, Daniel and Dingler, Felix A and Moignard, Victoria and Göttgens, Berthold and others (2014). Single-cell rna sequencing reveals t helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell reports*, *7*(4), 1130–1142.

Mahmud, Shawn A and Manlove, Luke S and Farrar, Michael A (2013). Interleukin-2 and stat5 in regulatory t cell development and function. *Jak-stat*, *2*(1), e23154.

M.K. El Mahrsi and E. Côme and J. Baro and L. Oukhellou (2014). Understanding passenger patterns in public transit through smart card and socioeconomic data a case study in rennes, france. *3rd International Workshop on Urban Computing (SigKDD)*.

Mankoo, Parminder K and Shen, Ronglai and Schultz, Nikolaus and Levine, Douglas A and Sander, Chris (2011). Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One*, *6*(11), e24709.

Martella, Francesca and Alfo, Marco and Vichi, Maurizio (2008). Biclustering of gene expression data by an extension of mixtures of factor analyzers. *The International Journal of Biostatistics*, *4*(1), 1–19.

Maynard, Nathaniel D and Chen, Jing and Stuart, Rhona K and Fan, Jian-Bing and Ren, Bing (2008). Genome-wide mapping of allele-specific protein-dna interactions in human cells. *Nature Methods*, *5*(4), 307–309.

G.J. McLachlan and K.-A. Do and C. Ambroise (2004). *Finite Mixture Models*. Wiley, New York.

Miller, David J and Wang, Yue and Kesidis, George (2008). Emergent unsupervised clustering paradigms with potential application to bioinformatics. *Frontiers in Bioscience: A Journal and Virtual Library*, *13*, 677–690.

Milligan, Glenn W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*(3), 325–342.

Miyagi, Atsuko and Takahashi, Hideyuki and Takahara, Kentaro and Hirabayashi, Takayuki and Nishimura, Yoshiki and Tezuka, Takafumi and Kawai-Yamada, Maki and Uchimiya, Hirofumi (2010). Principal component and hierarchical clustering analysis of metabolites in destructive weeds ; polygonaceous plants. *Metabolomics*, *6*(1), 146–155.

Pablo Montero and José A. Vilar (2014). TSclust An R Package for Time Series Clustering. *Journal of Statistical Software*, *62*(1), 1–43.

Morency, Catherine and Trépanier, Martin and Agard, Bruno (2006). Analysing the variability of transit users behaviour with smart card data. *Intelligent Transportation Systems Conference, 2006. ITSC'06.* IEEE, 44–49.

Morency, Catherine and Trépanier, Martin and Piché, Daniel and Chapleau, Robert (2010). Bridging the gap between complex data and decision-makers an example of an innovative interactive tool. *Transportation Planning and Technology*, *33*(6), 465–479.

Murray, Paul W. and Agard, Bruno and Barajas, Marco A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. *Computers and Industrial Engineering*, *109*, 233 – 252.

Murtagh, F. and Contreras, P. (2011). Methods of Hierarchical Clustering. *ArXiv e-prints arXiv:1105.0121*.

Murtagh, Fionn and Legendre, Pierre (2014). Ward's hierarchical agglomerative clustering method which algorithms implement ward's criterion ? *Journal of Classification*, *31*(3), 274–295.

Nantes, Alfredo and Ngoduy, Dong and Bhaskar, Ashish and Miska, Marc and Chung, Edward (2016). Real-time traffic state estimation in urban corridors from heterogeneous data. *Transportation Research Part C: Emerging Technologies*, *66*, 99–118.

Jordi Nin and David Carrera and Daniel Villatoro (2013). On the use of social trajectory-based clustering methods for public transport optimization. *Citizen in Sensor Networks - Second International Workshop, CitiSens2013, Barcelona, Spain, September 19, 2013, Revised Selected Papers*. 59–70.

Norris, Jeremy L and Farrow, Melissa A and Gutierrez, Danielle B and Palmer, Lauren D and Muszynski, Nicole and Sherrod, Stacy D and Pino, James C and Allen, Jamie L and Spraggins, Jeffrey M and Lubbock, Alex LR and others (2017). Integrated, high-throughput, multiomics platform enables data-driven construction of cellular responses and reveals global drug mechanisms of action. *Journal of Proteome Research*, *16*(3), 1364–1375.

Nugent, Rebecca and Meila, Marina (2010). An overview of clustering applied to molecular biology, 369–404.

Oh, Man-Suk and Raftery, Adrian E (2007). Model-based clustering with dissimilarities: A bayesian approach. *Journal of Computational and Graphical Statistics*, *16*(3), 559–585.

Ortega-Tong, Meisy A. (2013). *Classification of London's public transport users using smart card data*. master thesis, Massachusetts Institute of Technology. Department of Civil and Environmental Engineering.

Orzechowski, Patryk and Boryczko, Krzysztof (2016). *Text Mining with Hybrid Biclustering Algorithms*, Springer International Publishing, Cham. 102–113.

Owens, Clifford Conley (2009). *Mining truth tables and straddling biclusters in binary datasets*. master thesis, Virginia Polytechnic Institute and State University.

Pal, Nikhil R and Pal, Sankar K (1993). A review on image segmentation techniques. *Pattern Recognition*, *26*(9), 1277–1294.

Pan, Wenying and Ngo, Thuy TM and Camunas-Soler, Joan and Song, Chun-Xiao and Kowarsky, Mark and Blumenfeld, Yair J and Wong, Ronald J and Shaw, Gary M and Stevenson, David K and Quake, Stephen R (2016). Simultaneously monitoring immune response and microbial infections during pregnancy through plasma cfrna sequencing. *Clinical Chemistry*, clinchem–2017.

Park, Jin and Kim, Dong-Jun and Lim, Yongtaek (2008). Use of smart card data to define public transit use in seoul, south korea. *Transportation Research Record: Journal of the Transportation Research Board*, (2063), 3–9.

Partovi Nia, Vahid and Davison, Anthony C (2015). A simple model-based approach to variable selection in classification and clustering. *Canadian Journal of Statistics*, *43*(2), 157–175.

Partovi Nia, Vahid and Lysy, MArtin AND Mouret, Geoffroy (2017). No-means clustering: A stochastic variant of k-means. technical report, Les Cahiers du GERAD G-2017-33.

Ashish Kumar Patnaik and Prasanta Kumar Bhuyan and K.V. Krishna Rao (2016). Divisive analysis (diana) of hierarchical clustering and gps data for level of service criteria of urban streets. *Alexandria Engineering Journal*, *55*(1), 407 – 418.

Pelletier, Marie-Pier and Trépanier, Martin and Morency, Catherine (2011). Smart card data use in public transit a literature review. *Transportation Research Part C Emerging Technologies*, *19*(4), 557–568.

Beatriz Pontes and Raúl Giráldez and Jesús S. Aguilar-Ruiz (2015). Biclustering on expression data: A review. *Journal of Biomedical Informatics*, *57*, 163 – 180.

Punj, Girish and Stewart, David W (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 134–148.

Rathipriya, R and Thangavel, K (2014). Extraction of web usage profiles using simulated annealing based biclustering approach. *arXiv preprint arXiv:1412.8099*.

Alvin C. Rencher (1998). *Multivariate Statistical Inference and Applications*. Wiley, New York.

Ritchie, Marylyn D and Holzinger, Emily R and Li, Ruowang and Pendergrass, Sarah A and Kim, Dokyoon (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, *16*(2), 85.

Romero, Roberto and Erez, Offer and Maymon, Eli and Chaemsaithong, Piya and Xu, Zhonghui and Pacora, Percy and Chaiworapongsa, Tinnakorn and Done, Bogdan and Hassan, Sonia S and Tarca, Adi L (2017). The maternal plasma proteome changes as a function of gestational age in normal pregnancy: a longitudinal study. *American Journal of Obstetrics and Gynecology*, *217*, 61.e1, 61.e21.

Rugeles, Daniel and Zhao, Kaiqi and Gao, Cong and Dash, Manoranjan and Krishnaswamy, Shonali (2017). Biclustering: An application of dual topic models. *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 453–461.

Schadt, Eric E and Lamb, John and Yang, Xia and Zhu, Jun and Edwards, Steve and GuhaThakurta, Debraj and Sieberts, Solveig K and Monks, Stephanie and Reitman, Marc and Zhang, Chunsheng and others (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, *37*(7), 710–717.

Schrage, LE (1967). The queue m/g/1 with feedback to lower priority queues. *Management Science*, *13*(7), 466–474.

Schwarz, Gideon E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Schwenk, Jochen M and Igel, Ulrika and Kato, Bernet S and Nicholson, George and Karpe, Fredrik and Uhlén, Mathias and Nilsson, Peter (2010). Comparative protein profiling of serum and plasma using an antibody suspension bead array approach. *Proteomics*, *10*(3), 532–540.

Serra, Angela and Fratello, Michele and Fortino, Vittorio and Raiconi, Giancarlo and Tagliaferri, Roberto and Greco, Dario (2015). Mvda: a multi-view genomic data integration methodology. *BMC Bioinformatics*, *16*(1), 261.

Shabalin, Andrey A (2012). Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, *28*(10), 1353–1358.

Sharkey, Amanda J C (1996). On combining artificial neural nets. *Connection Science*, *8*(3-4), 299–314.

Shekhar, Shashi and Jiang, Zhe and Ali, Reem Y. and Eftelioglu, Emre and Tang, Xun and Gunturi, Venkata M. V. and Zhou, Xun (2015). Spatiotemporal data mining a computational perspective. *ISPRS International Journal of Geo-Information*, *4*(4), 2306–2338.

Shen, Ronglai and Olshen, Adam B and Ladanyi, Marc (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*(22), 2906–2912.

Sheng, Qizheng and Moreau, Yves and De Moor, Bart (2003). Biclustering microarray data by Gibbs sampling. *Bioinformatics*, *19*, 196–205.

Wang Shirui (2016). *Spatiotemporal Visual Analysis of Traffi Flow Patterns Related to Transport Hubs from Floating Car Data*. master thesis, Technische Universität München.

Smith, J.Q. and Anderson, P.E. and Liverani, S. (2008). Separation measures and the geometry of Bayes factor selection for classification. *Journal of the Royal Statistical Society, Series B*, *70*, 957–980.

Sneath, P. H. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, *17*(1), 201–226. First algorithm of hierarchical clustering.

Sokal, Robert R (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.

R. R. Sokal and P. H. Sneath (1963). *Principles of Numerical Taxonomy*. Freeman, London.

Sørensen, Thorvald (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, *5*, 1–34. First complete Linkage suggestion.

Spearman, Charles (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Sreekumar, Arun and Poisson, Laila M and Rajendiran, Thekkelnaycke M and Khan, Amjad P and Cao, Qi and Yu, Jindan and Laxman, Bharathi and Mehra, Rohit and Lonigro, Robert J and Li, Yong and others (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, *457*(7231), 910.

Stahl, Daniel and Sallis, Hannah (2012). Model-based cluster analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(4), 341–358.

Stevenson, DK and Shaw, GM and Wise, PH and Norton, ME and Druzin, ML and Valantine, HA and McFarland, DA and March of Dimes Prematurity Research Center at Stanford University School of Medicine and others (2013). Transdisciplinary translational science and the case of preterm birth. *Journal of Perinatology*, *33*(4), 251.

Sun, Rui and Li, Ai Ling and Wei, Hai Ming and Tian, Zhi Gang (2004). Expression of prolactin receptor and response to prolactin stimulation of human nk cell lines. *Cell Research*, *14*(1), 67.

Tan, Pang-Ning and others (2006). *Introduction to data mining*. Pearson Education India.

Telgarsky, Matus and Dasgupta, Sanjoy (2012). Agglomerative bregman clustering. *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 1011–1018.

Tenenhaus, Arthur and Tenenhaus, Michel (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Research*, *238*(2), 391–403.

Tibshirani, Robert (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tibshirani, Robert and Walther, Guenther and Hastie, Trevor (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B*, *63*(2), 411–423.

Martin Trépanier and Nicolas Tranchant and Robert Chapleau (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, *11*(1), 1–14.

Tu, Kewei and Honavar, Vasant (2008). Unsupervised learning of probabilistic context-free grammar using iterative biclustering. *International Colloquium on Grammatical Inference*. Springer, 224–237.

Michele Tumminello and Fabrizio Lillo and Rosario N. Mantegna (2010). Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior and Organization*, *75*(1), 40 – 58. Transdisciplinary Perspectives on Economic Complexity.

Heather Turner and Trevor Baileya and Wojtek Krzanowski (2005). Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis*, *48*, 235–254.

Tyanova, Stefka and Temu, Tikira and Sinitcyn, Pavel and Carlson, Arthur and Hein, Marco Y and Geiger, Tamar and Mann, Matthias and Cox, Jürgen (2016). The perseus computational platform for comprehensive analysis of (prote) omics data. *Nature Methods*, *13*(9), 731–740.

van Uitert, M. and Meuleman, W. and Wessels, L. (2008). Biclustering sparse binary genomic data. *Journal of Computational Biology*, *15*, 1329–1345.

Van Wijk, Jarke J. and Van Selow, Edward R. (1999). Cluster and calendar based visualization of time series data. *Proceedings of the 1999 IEEE Symposium on Information Visualization*. IEEE Computer Society, Washington, DC, USA, INFOVIS '99, 4–9.

Ulrike von Luxburg and Robert C. Williamson and Isabelle Guyon (2012). Clustering: Science or art ? I. Guyon, G. Dror, V. Lemaire, G. Taylor and D. Silver, editors, *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. PMLR, Bellevue, Washington, USA, vol. 27 of *Proceedings of Machine Learning Research*, 65–79.

Jonas De Vos and Frank Witlox (2013). Transportation policy as spatial planning tool ; reducing urban sprawl by increasing travel costs and clustering infrastructure and public transportation. *Journal of Transport Geography*, *33*(Supplement C), 117 – 125.

Walsh, Scott TR and Kossiakoff, Anthony A (2006). Crystal structure and site 1 binding energetics of human placental lactogen. *Journal of Molecular Biology*, *358*(3), 773–784.

J. Wang and H. Song and X. Zhou (2015). A collaborative filtering recommendation algorithm based on biclustering. *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. 803–807.

J. Wang and W. Yu (2010). Research on hierarchy of urban rail transit hub. *2010 International Conference on Intelligent Computation Technology and Automation*. vol. 2, 1173–1176.

Wang, Shuang-Quan and Yang, Jie and Chou, Kuo-Chen (2006). Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. *Journal of Theoretical Biology*, *242*(4), 941–946.

Yu-Xiong Wang and Yu-Jin Zhang (2013). Nonnegative matrix factorization a comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, *25*(6), 1336–1353.

Weisbrod, Glen and Reno, Arlee (2009). *Economic impact of public transportation investment*. American Public Transportation Association.

Witten, Daniela M and Tibshirani, Robert (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, *105*(490), 713–726.

Wolpert, David H (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259.

Xu, Bin and Bu, Jiajun and Chen, Chun and Cai, Deng (2012). An exploration of improving collaborative recommender systems via user-item subgroups. *Proceedings of the 21st International Conference on World Wide Web*. ACM, 21–30.

Q Xu and B H Mao and Y Bai (2016). Network structure of subway passenger flows. *Journal of Statistical Mechanics: Theory and Experiment*, *2016*(3), Article ID 033404.

Xu, Rui and Wunsch, Donald (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645–678.

Yahya, Saadiah and Noor, Noriani Mohammed (2008). Strategic planning of an integrated smart card fare collection system - challenges and solutions. *Proceedings of the 2008 11th IEEE International Conference on Computational Science and Engineering*. Washington, DC, USA, 31–36.

Yang, Pengyi and Hwa Yang, Yee and B Zhou, Bing and Y Zomaya, Albert (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, *5*(4), 296–308.

Yau, Christopher and others (2016). pcareduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, *17*(1), 140.

Yeung, K.Y. and Medvedovic, M. and Bumgarner, R.E. (2003). Clustering gene-expression data with repeated measurements. *Genome Biology*, *4*, R34.

Yeung, K.Y. and Ruzzo, W.L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, *17*, 763–774.

Zamir, Oren and Etzioni, Oren (1998). Web document clustering: A feasibility demonstration. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 46–54.

Zhang, J. (2010). A Bayesian model for biclustering with applications. *Journal of the Royal Statistical Society, Series C*, *59*, 635–656.

Zhao, Xing-Ming and Chen, Luonan and Aihara, Kazuyuki (2008). Protein function prediction with high-throughput data. *Amino Acids*, *35*(3), 517.

Zhu, Jun and Sova, Pavel and Xu, Qiuwei and Dombek, Kenneth M and Xu, Ethan Y and Vu, Heather and Tu, Zhidong and Brem, Rachel B and Bumgarner, Roger E and Schadt, Eric E (2012). Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biology*, *10*(4), e1001301.

Zhu, Jun and Zhang, Bin and Smith, Erin N and Drees, Becky and Brem, Rachel B and Kruglyak, Leonid and Bumgarner, Roger E and Schadt, Eric E (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, *40*(7), 854–861.

Zou, Hui and Hastie, Trevor (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320.